TRIANGULAR FINITE ELEMENT SOLUTION TO BOUNDARY VALUE PROBLEMS

A thesis submitted to
Lakehead University
in partial fulfillment of the requirements
for the degree of
Master of Science

by

ONG HOON LIONG

1977

ACKNOWLEDGEMENT

I wish to thank my supervisor, Professor P.O. Frederickson, for his advice and encouragement during the preparation of this thesis.

ABSTRACT

This Thesis discusses the triangular finite element solution to second order elliptic boundary value problems. The Barycentric Coordinate system, which some engineers call the areal coordinate system, is used throughout in this Thesis. Some fundamental parts of vector calculus are developed in this coordinate system, and are applied to the triangular finite element method.

We also present a new approach to error analysis based on the computation of Peano-Sard kernels [F6] of error functionals in the Barycentric Coordinate system. Some numerical quadrature formulas for the approximation of the load vector $F^h = \int f\phi \, d\mu$ are derived, and error bounds are estimated.

Several approximate inversion methods for the construction of an ε-approximate inverse to A in the iterative solution of the linear system $Ax = y$ are discussed. These procedures include the truncation (TRq) method [B3], the least-squares (LSq) method [B3], the weighted truncation (WTq) method and the interpolation (INq) method. These ε-approximate inverses are applied to the iterative algorithm FAPIN [F4] to solve the linear system $Ax = y$.

To illustrate the theory, three boundary value problems are solved numerically using piecewise linear splines in the Ritz-Galerkin method. Inhomogeneous boundary conditions are used in two

of the problems, and in one of these the differential operator is singular.

CONTENTS

CHAPTER 1

FUNDAMENTAL CONCEPTS

1.1 INTRODUCTION

We begin this chapter by introducing the Barycentric Coordinate system, which some engineers call the areal coordinate system, and which is essential for a study of the triangular finite element method. Some fundamental results are given in this chapter that serve as a basis for the chapters that follow.

Sobolev spaces and Sobolev norms are defined in terms of Barycentric Coordinates in Section 1.6. We state the generalized Peano-Sard Kernel Theorem in Section 1.7, followed by an example on the application of the Theorem and the construction of kernels of the error functional $E(f)$. Further demonstration on the application of this Theorem will be given in Chapter 3. The non-uniqueness of the kernel is shown by giving an example.

1.2 BASIC NOTATION

Definition 1.2.1. Let τ be a set of triangles in a bounded polygonal domain Ω. We say τ is a triangulation of Ω ([S4], [P1], [B6]) if

(i) for each pair of distinct triangles in τ, they either intersect at exactly one vertex or intersect on one complete side or do not intersect at all.

(ii) the union of all the triangles in τ and their interior is Ω.

We will denote by τ^h the triangulation of Ω, such that each element of τ^h is an equilateral triangle of side length equal to h. We also denote by Ω_h the set of all vertices of triangles T in τ^h. Elements of Ω_h are called *nodes* of τ^h. A node of τ^h is called an *interior node* if it does not lie on the boundary $\partial\Omega$ of Ω. The set of all interior nodes will be denoted by $\mathring{\Omega}_h$.

Let L be the integer lattice in the plane. Since every element of L can be written as a linear combination of $(1,0)$, $(-1,1)$, $(0,-1)$ over the set of integer N, we can define a norm, called the *hexagonal norm* on L by

$$|\alpha| = \min \{ \sum_{j=1}^{3} |k_j| : \alpha = k_1(1,0)+k_2(-1,1)+k_3(0,-1), k_j \in N \}.$$

For each triangulation τ^h of Ω, there is an 1-1 correspondence between the set Ω_h and a subset Γ_h of L with the property that : for every T in τ^h, the distance between any two of the corresponding vertices of T in Γ_h is one. Elements of Ω_h will be denoted by X_α, where $\alpha \in \Gamma_h$. We denote by $\mathring{\Gamma}_h$ the set of $\alpha \in \mathring{\Gamma}_h$ s.t. $X_\alpha \in \mathring{\Omega}_h$.

We observe that for every member X_α of $\mathring{\Omega}_h$, the set $\{ X_\beta \in \Omega_h : |\alpha-\beta| = 1 \}$ form a hexagon in Ω with centre X_α; this hexagon will be denoted by $X_\alpha + H$.

Denote by $P^n(\Omega)$ the space of polynomials of degree $\leq n$, by $C(\Omega)$ the space of all continuous real valued functions defined on Ω, and by $C^n(\Omega)$ the space of real valued functions with continuous derivatives of order up to n.

Denote by $S^{n,q}$ the class of q-times differentiable real valued functions which are piecewise polynomials of degree n in each of the triangular elements $T \in \tau$. In particular, we will refer to the elements of $S^{1,0}$ as *linear splines* [F3].

Definition 1.2.2. A subset Z of a linear space X is an *affine space* iff $\lambda x_1 + (1-\lambda)x_2 \in Z$ whenever $x_1, x_2 \in Z$ and $\lambda \in R$. A function $\xi : X \to X$ is *affine* iff $\xi[\lambda x_1+(1-\lambda)x_2] = \lambda\xi(x_1) + (1-\lambda)\xi(x_2)$ whenever $x_1, x_2 \in X$ and $\lambda \in R$.

Remark : Every affine function can be written as a linear function plus a constant C, thus, an affine function is linear iff the constant C is zero.

1.3 BARYCENTRIC COORDINATES

Let $T = A_0A_1A_2$ be any triangular element in τ. Consider the affine space X, generated by the three vertices of T, i.e.

$$X = \{ \sum_i \xi_i A_i : \sum_i \xi_i = 1 \}.$$

Since A_0, A_1, A_2 are not collinear, they are affinely independent, and any point $P \in X$ can be uniquely represented as

$$P = \xi_0 A_0 + \xi_1 A_1 + \xi_2 A_2 \qquad \xi_0 + \xi_1 + \xi_2 = 1 \qquad (1.3.1)$$

Denote by ξ_i, $i = 0, 1, 2$ the three affine functions defined by the equation $\xi_i(A_j) = \delta_{i,j}$, where δ denotes the Kronecker delta function ([F3], [L1]). Then, we have

$$\xi_i(P) = \xi_i \quad \text{for } i = 0, 1, 2 \tag{1.3.2}$$

Since the expression (1.3.1) is unique, the point P can be represented as

$$P = (\xi_0(P), \xi_1(P), \xi_2(P)) \quad \text{or} \quad \text{simply } P(\xi_0, \xi_1, \xi_2).$$

We will refer to this as the Barycentric Coordinates of P w.r.t. the triangle $T = A_0A_1A_2$.

If P is in the interior of T, then we have $0 < \xi_i < 1$, $i = 0, 1, 2$. Geometrically ([S4], [H4]), ξ_i is the ratio of

$$\frac{\text{Length of } PQ}{\text{Length of } A_iQ} = \frac{\text{Area of } PA_{i+1}A_{i-1}}{\text{Area of } A_iA_{i+1}A_{i-1}}$$

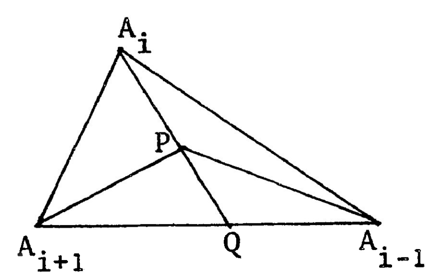


Fig. 1.3.1

Where Q is the point of intersection of the two straight lines A_iP and $A_{i+1}A_{i-1}$ as shown in Fig. 1.3.1. For this reason, some authors refer to this as an Areal Coordinate system. We might also use the term Affine Coordinate system.

We observe that $\xi_i(P)$ remains unchanged for all P lying on a line parallel to the side $A_{i+1}A_{i-1}$, in particular

$$\xi_i(A_{i+1}A_{i-1}) = 0.$$

Any polynomial of degree n on Ω can be expressed uniquely as a homogeneous polynomial of degree n in the Barycentric Coordinates w.r.t. a specific triangle $T = A_0A_1A_2$ in τ, or else as a polynomial of degree n in any two of the coordinates. For example, the polynomial $\xi_0\xi_1\xi_2$ which vanishes on all three sides of the triangle T is equivalent to the inhomogeneous polynomial $\xi_1\xi_2 - \xi_1^2\xi_2 - \xi_1\xi_2^2$.

In order to compute the integral of polynomials over individual triangular element in τ, it is convenient to express the polynomial in terms of Barycentric Coordinates locally. But then another problem arises : the same polynomial will have different local expression over each of the triangular elements. This problem can be solved by establishing the relationships between the Barycentric Coordinates of a point $X \in \Omega$ w.r.t. two different triangles in τ.

Let $T_A = A_0A_1A_2$ and $T_B = B_0B_1B_2$ be any two triangular elements in τ. Suppose (ξ_0,ξ_1,ξ_2) and (η_0,η_1,η_2) are the Barycentric Coordinates of a point $X \in \Omega$ w.r.t. T_A and T_B respectively. It follows from (1.3.1) and (1.3.2) that X can be represented as :

$$X = \xi_0A_0 + \xi_1A_1 + \xi_2A_2 = \eta_0B_0 + \eta_1B_1 + \eta_2B_2 \qquad (1.3.3)$$

Denote by $(\xi_0^{B_i},\xi_1^{B_i},\xi_2^{B_i})$, $i = 0, 1, 2$ the Barycentric

Coordinates of B_i w.r.t. T_A.

Then we have

$$B_i = \xi_0^{B_i} A_0 + \xi_1^{B_i} A_1 + \xi_2^{B_i} A_2 \qquad i = 0,\ 1,\ 2.$$

Substituting these into (1.3.3), we have

$$X = \sum_i \xi_i A_i = \sum_j \eta_j \left(\sum_i \xi_i^{B_j} A_i\right) = \sum_i \left(\sum_j \eta_j \xi_i^{B_j}\right) A_i$$

Since the representation of X in terms of A_0, A_1, A_2 is unique, we have

$$\xi_i = \sum_j \eta_j \xi_i^{B_j} \qquad i = 0,\ 1,\ 2 \tag{1.3.4}$$

It is plain that the transformation $\Phi(\eta_0,\eta_1,\eta_2) = (\xi_0,\xi_1,\xi_2)$ described by (1.3.4) is a linear transformation; the map Φ can be written in matrix form as :

$$\tilde{\Phi} = \begin{pmatrix} \xi_0^{B_0} & \xi_0^{B_1} & \xi_0^{B_2} \\ \xi_1^{B_0} & \xi_1^{B_1} & \xi_1^{B_2} \\ \xi_2^{B_0} & \xi_2^{B_1} & \xi_2^{B_2} \end{pmatrix} \tag{1.3.5}$$

and $$\Phi(\eta_0,\eta_1,\eta_2) = \left(\tilde{\Phi}\begin{pmatrix}\eta_0\\ \eta_1\\ \eta_2\end{pmatrix}\right)^T$$

Thus, if f is a function mapping Ω into R, then by virtue of (1.3.1) and (1.3.2), there exists a function

$$F : \mathcal{R}^3 \to \mathcal{R} \quad \text{s.t.}$$

$$f(P) = F(\xi_0(P), \xi_1(P), \xi_2(P))^{\dagger} \qquad (1.3.6)$$

i.e. the function f can be expressed as a function F in terms of the Barycentric Coordinates of P w.r.t. T_A. It follows from (1.3.4) that the function f can also be expressed as a function in terms of the Barycentric Coordinates of P w.r.t. T_B as

$$f(P) = F \cdot \Phi(\eta_0(P), \eta_1(P), \eta_2(P)) \qquad (1.3.7)$$

where Φ is the linear transformation characterized by the matrix $\tilde{\Phi}$ given in (1.3.5).

In particular, we are interested to look at the six matrices $\tilde{\Phi}_j$ of the hexagon $A_\alpha + H$. As shown in Fig. 1.3.2, let $T_j = A_\alpha A_{\alpha_j} A_{\alpha_{j+1}}$ be the six triangles of the hexagon $A_\alpha + H$. If $F(\xi_0, \xi_1, \xi_2)$ is an expression of a function $f : \Omega \to \mathcal{R}$ in terms of the Barycentric Coordinates w.r.t. T_1, then, $F \cdot \Phi_j(\xi_0, \xi_1, \xi_2)'$ is an expression of f in terms of the Barycentric Coordinates w.r.t. T_j. The six linear transformation Φ_j are given by :

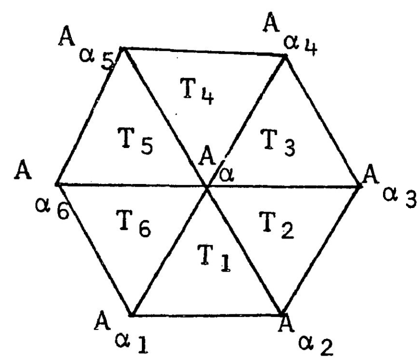


Fig. 1.3.2

$$\tilde{\Phi}_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

† This representation F is not unique.

$$\tilde{\Phi}_2 = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 0 & -1 \\ 0 & 1 & 1 \end{pmatrix}$$

$$\tilde{\Phi}_3 = \begin{pmatrix} 1 & 1 & 2 \\ 0 & -1 & -1 \\ 0 & 1 & 0 \end{pmatrix}$$

$$\tilde{\Phi}_4 = \begin{pmatrix} 1 & 2 & 2 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

$$\tilde{\Phi}_5 = \begin{pmatrix} 1 & 2 & 1 \\ 0 & 0 & 1 \\ 0 & -1 & -1 \end{pmatrix}$$

$$\tilde{\Phi}_6 = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & -1 & 0 \end{pmatrix}$$

1.4 DIFFERENTIATION AND INTEGRATION IN BARYCENTRIC COORDINATES

Let $T = A_0A_1A_2$ be a triangle in Ω. Define the first order linear differential operator w.r.t. the Barycentric Coordinate ξ_i [F6] by $D_i(\xi_i) = 0$ and $D_i(\xi_{i\pm 1}) = \mp 1$ i.e. the counter clockwise normalized derivative of a function f in the direction parallel to the opposite side of A_i.

If f is a function mapping Ω into $\mathcal{R}$, and if

$F(\xi_0,\xi_1,\xi_2)$ is an expression of f w.r.t. the triangle T, then we have

$$D_i f = \frac{\partial F}{\partial \xi_{i-1}} - \frac{\partial F}{\partial \xi_{i+1}} \tag{1.4.1}$$

Let f and g be two real valued functions defined on Ω. If the derivatives $D_i f$ and $D_i g$ exist, then the operator D_i has the following properties :

1. $D_i(f+g) = D_i f + D_i g$
2. $D_i(cf) = cD_i f$ for any constant $c \in R$
3. $\sum_i D_i f = 0$ (1.4.2)
4. $D_i(gf) = gD_i f + fD_i g$ (1.4.3)
5. $D_i(\frac{f}{g}) = \frac{D_i f}{g} - \frac{f}{g^2} D_i g$ if $g \neq 0$

The differential operator can be extended to any order through $D^\alpha f = D_0^{i_1} D_1^{i_2} D_2^{i_3} f$, where $\alpha = (i_1,i_2,i_3) \in N^3$, D^0 and D_i^0 denote the *identity operator*. We will denote by $|\alpha| = i_1 + i_2 + i_3$ the order of derivative of f.

Define $\int_E f \, d\mu_T$ the normalized Lebesgue integral of f on a measurable subset E of T, s.t. $\int_T 1 \, d\mu_T = 1$.

Define $\int_E f \, d\mu_\Omega$ the normalized Lebesgue integral of f on a measurable subset E of Ω, s.t. $\int_\Omega 1 \, d\mu_\Omega = 1$.

If Ω is a bounded polygonal region and τ is a triangulation of Ω, then we have

$$\int_{\Omega} f \, d\mu_{\Omega} = \sum_{T \in \tau} \mu_{\Omega}(T) \int_{T} f \, d\mu_{T}$$

In particular, if Ω is an equilateral triangle of unit side length and τ^h is an equilateral triangulation of Ω, we have $\mu_{\Omega}(T) = h^2$ for all $T \in \tau^h$. i.e.

$$\int_{\Omega} f \, d\mu_{\Omega} = h^2 \sum_{T \in \tau^h} \int_{T} f \, d\mu_{T}$$

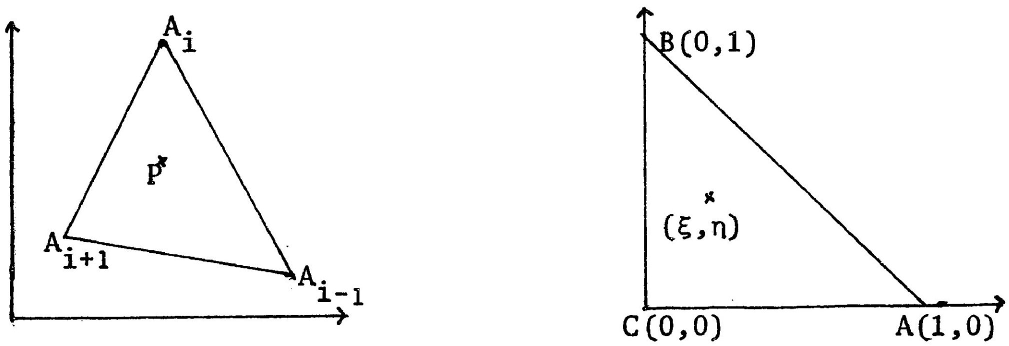


Fig. 1.4.1

As shown in Fig. 1.4.1, the triangle $T = A_i A_{i+1} A_{i-1}$ can be transformed into the standard triangle A(1,0), B(0,1), C(0,0) by using the affine function which maps $A_i \to C$, $A_{i+1} \to A$, and $A_{i-1} \to B$. Thus if $f : T \to \mathbb{R}$, then there exists a function

$$F : ABC \to \mathbb{R} \text{ s.t.}$$

$$f(P) = F(\xi,\eta) \tag{1.4.4}$$

where (ξ,η) is the affine image of the point $P \in T$. The

Jacobian of this transformation is 2. Thus, the integral $\int_T f \, d\mu_T$ can also be written as :

$$\int_T f \, d\mu_T = 2\iint_{ABC} F \, d\xi d\eta = 2\int_0^1\int_0^{1-\xi} F \, d\xi d\eta \qquad (1.4.5)$$

Define $\int_{A_{i+1}}^{A_{i-1}} f(X) \, dX$ as the Lebesgue line integral of f along the line $A_{i+1}A_{i-1}$, normalized by $\int_{A_{i+1}}^{A_{i-1}} 1 \, dX = 1$.

The following lemmas are some important properties of line and surface integrals :

Lemma 1.4.1. Let $f : T \to R$, if $D_i f$ exists on the side $A_{i+1}A_{i-1}$, then $\int_{A_{i+1}}^{A_{i-1}} D_i f(X) \, dX = f(A_{i-1}) - f(A_{i+1})$ (1.4.6)

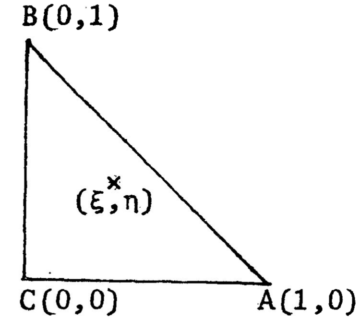
Proof : Using the affine transformation to map $A_{i+1}A_{i-1}$ onto $[0,1]$ through $A_{i+1} \to 0$ and $A_{i-1} \to 1$, then there exists a function $F : [0,1] \to R$, s.t. $f(P) = F(t)$ where t is the affine image of P. Thus,

$$\int_{A_{i+1}}^{A_{i-1}} D_i f(X) \, dX = \int_0^1 F'(t) \, dt = F(1) - F(0) = f(A_{i-1}) - f(A_{i+1})$$

Lemma 1.4.2. $\int_{A_{i+1}}^{A_{i-1}} gD_i f \, dX = (gf)(A_{i-1}) - (gf)(A_{i+1}) - \int_{A_{i+1}}^{A_{i-1}} fD_i g \, dX$ (1.4.7)

Proof : The result follows from (1.4.3) and Lemma 1.4.1.

Lemma 1.4.3. $\int_T D_i f \, d\mu_T = 2\int_{A_{i-1}}^{A_i} f \, dX - 2\int_{A_i}^{A_{i+1}} f \, dX$ (1.4.8)



Proof : From (1.4.1) and (1.4.5), we have

$$\int_T D_i f \, d\mu_T = 2\iint_{ABC} \left(\frac{\partial F}{\partial \eta} - \frac{\partial F}{\partial \xi}\right) d\xi d\eta$$

By Green's Theorem [H2], we get

$$\int_T D_i f \, d\mu_T = -2\left(\oint F \, d\xi + \oint F \, d\eta\right)$$

The symbol $\oint$ denotes the line integral along the three sides of the triangle CAB in the counter clockwise direction.

Since $\xi + \eta = 1$ for all point $P(\xi,\eta)$ on AB, we have

$$\int_A^B F \, d\xi + \int_A^B F \, d\eta = \int_A^B F \, d(\xi+\eta) = 0$$

It follows that

$$\int_T D_i f \, d\mu_T = -2\left(\int_C^A F \, d\xi + \int_B^C F \, d\xi + \int_C^A F \, d\eta + \int_B^C F \, d\eta\right)$$

$$= -2\left(\int_C^A F \, d\xi + \int_B^C F \, d\eta\right)$$

Since $\int_{A_i}^{A_{i+1}} f \, dX = \int_C^A F \, d\xi$ and $\int_{A_{i-1}}^{A_i} f \, dX = -\int_B^C F \, d\eta$, we get

$$\int_T D_i f \, d\mu_T = 2\int_{A_{i-1}}^{A_i} f \, dX - 2\int_{A_i}^{A_{i+1}} f \, dX.$$

Lemma 1.4.4. $\int_T gD_i f \, d\mu_T = 2\int_{A_{i-1}}^{A_i} fg \, dX - 2\int_{A_i}^{A_{i+1}} fg \, dX - \int_T fD_i g \, d\mu_T$ (1.4.9)

Proof : The result follows from (1.4.3) and Lemma 1.4.3.

We shall end this section by stating two very useful formulas of line and surface integrals of polynomials of the form : $\xi_1^{s_1}\xi_2^{s_2}\xi_3^{s_3}$, where s_i, $i = 1, 2, 3$ are three non-negative integers.

<u>Lemma 1.4.5.</u> $$\int_{A_{i+1}}^{A_{i-1}} \xi_1^{s_1}\xi_2^{s_2}\xi_3^{s_3}\, dX = \frac{s_1!s_2!s_3!}{(s_1+s_2+s_3+1)!}\,\delta_{0,s_i} \qquad (1.4.10)$$

where δ is the Kronecker delta function.

<u>Proof</u> : If $s_i \neq 0$ then the function $\xi_1^{s_1}\xi_2^{s_2}\xi_3^{s_3}$ vanishes on the side $A_{i+1}A_{i-1}$ and hence the right hand side of (1.4.10) vanishes. If $s_i = 0$, then

$$\int_{A_{i+1}}^{A_{i-1}} \xi_1^{s_1}\xi_2^{s_2}\xi_3^{s_3}\, dX = \int_{A_{i+1}}^{A_{i-1}} \xi_{i-1}^{s_{i-1}}\xi_{i+1}^{s_{i+1}}\, dX$$

It follows from the affine transformation defined by $A_{i+1} \to 0$ and $A_{i-1} \to 1$ that

$$\int_{A_{i+1}}^{A_{i-1}} \xi_{i-1}^{s_{i-1}}\xi_{i+1}^{s_{i+1}}\, dX = \int_0^1 t^{s_{i-1}}(1-t)^{s_{i+1}}\, dt$$

Applying itegration by parts to the above integral, the result (1.4.10) follows.

<u>Lemma 1.4.6.</u> $$\int_T \xi_1^{s_1} \xi_2^{s_2} \xi_3^{s_3} \, d\mu_T = \frac{s_1! s_2! s_3!}{(s_1+s_2+s_3+2)!} \tag{1.4.11}$$

Holand and Bell [H4,p.84] and T. H. Lim [L1,p.24] have presented a proof of the Lemma.

1.5 THE DEL OPERATOR ∇ AND THE LAPLACIAN OPERATOR Δ

Suppose Ω is an open subet of R^2 and U is a real valued function on Ω. Let e be a unit vector in Ω, then the derivative of U at a point $x \in \Omega$ in the direction e is defined as the limit

$$\tilde{D}_e U(X) = \lim_{\varepsilon \to 0} \frac{U(X+\varepsilon e)-U(X)}{\varepsilon} \tag{1.5.1}$$

when the limit exists.

If $U \in C^1(\Omega)$, then there exists a vector function ([W1,P.159], [H3,P.374])

$$\nabla U : \Omega \to R^2 \quad \text{s.t.}$$

$$e \cdot \nabla U(X) = \tilde{D}_e U(X) \tag{1.5.2}$$

for all unit vectors e in Ω. The function ∇U is called the *gradient* of U.

If $U \in C^2(\Omega)$, then the operator Δ defined by $\Delta U = \nabla \cdot \nabla U$ is called the *Laplacian operator*.

Let X be a point in the triangular element $T \in \tau^h$. Denote by e_i the unit vector in the direction $A_{i+1}A_{i-1}$, then the vector ∇U can be written in terms of $e_{i\pm 1}$ as follows :

$\nabla U = \lambda_{i+1} e_{i+1} + \lambda_{i-1} e_{i-1}$

where $\lambda_{i\pm 1}$ are to be determined.

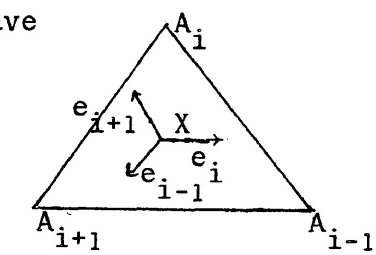
It follows from (1.5.2) that

$$\frac{1}{h} D_{i\pm 1} U = e_{i\pm 1} \cdot \nabla U = \lambda_{i+1} e_{i+1} \cdot e_{i\pm 1} + \lambda_{i-1} e_{i-1} \cdot e_{i\pm 1} \qquad (1.5.3)$$

Since $A_i A_{i+1} A_{i-1}$ is an equilateral triangle, we have

$$e_i \cdot e_j = \begin{cases} 1 & \text{if } i = j \\ -\frac{1}{2} & \text{if } i \neq j \end{cases} \qquad (1.5.4)$$

By substituting (1.5.4) into (1.5.3), we have



$$\begin{cases} \frac{1}{h} D_{i+1} U = \lambda_{i+1} - \frac{1}{2}\lambda_{i-1} \\ \frac{1}{h} D_{i-1} U = -\frac{1}{2}\lambda_{i+1} + \lambda_{i-1} \end{cases}$$

By solving the above linear equations, we obtain

$$\begin{cases} \lambda_{i+1} = \frac{2}{3h}(D_{i-1}U + 2D_{i+1}U) \\ \lambda_{i-1} = \frac{2}{3h}(D_{i+1}U + 2D_{i-1}U) \end{cases}$$

It follows that the gradient operator has a representation of the form :

$$\nabla = e_{i+1} \frac{2}{3h}(D_{i-1} + 2D_{i+1}) + e_{i-1} \frac{2}{3h}(D_{i+1} + 2D_{i-1})$$

$$= \frac{2}{3h}[e_{i+1}(D_{i+1} - D_i) + e_{i-1}(D_{i-1} - D_i)]$$

$$= \frac{2}{3h}[e_{i+1}D_{i+1} + e_{i-1}D_{i-1} - (e_{i+1} + e_{i-1})D_i]$$

Since $\sum_i e_i = 0$, we have a symmetric representation of ∇ as follow :

$$\nabla = \frac{2}{3h} \sum_i e_i D_i \tag{1.5.5}$$

Lemma 1.5.1. Let U be differentiable in an open subset Ω of R^2, then at each point X in Ω for which $\nabla U(X) \neq 0$, the vector $\nabla U(X)$ points in the direction in which the derivative of U is numerically greatest, and the number $|\nabla U(X)|$ is equal to that maximum derivative.

Proof : Let e be a unit vector at a point X in Ω for which $\nabla U(X) \neq 0$. By equation (1.5.2), we have

$$\tilde{D}_e U(X) = e \cdot \nabla U(X) = |\nabla U(X)| \cos\theta \leq |\nabla U(X)| \tag{1.5.6}$$

where θ is the angle between the two vectors e and $\nabla U(X)$. The inequality (1.5.6) is sharp iff

$$e = \frac{\nabla U(X)}{|\nabla U(X)|}$$

completing the proof.

Lemma 1.5.2. Let $T \in \tau^h$. If U and V are two differentiable functions in the triangle T, then

$$\nabla U \cdot \nabla V = \frac{2}{3h^2} \sum_i (D_i U)(D_i V) \tag{1.5.7}$$

Proof : It follows from equation (1.5.5) that

$$\begin{aligned}
\nabla U \cdot \nabla V &= \left(\frac{2}{3h} \sum_i e_i D_i U\right) \cdot \left(\frac{2}{3h} \sum_i e_i D_i V\right) \\
&= \frac{4}{9h^2} \sum_i [e_i \cdot e_i (D_i U)(D_i V) + e_i \cdot e_{i+1} (D_i U)(D_{i+1} V) + \\
&\qquad e_i \cdot e_{i-1} (D_i U)(D_{i-1} V)] \\
&= \frac{4}{9h^2} \sum_i [(D_i U)(D_i V) - \frac{1}{2}(D_i U)(D_{i+1} V + D_{i-1} V)] \\
&= \frac{4}{9h^2} \sum_i [(D_i U)(D_i V) + \frac{1}{2}(D_i U)(D_i V)] \\
&= \frac{4}{3h^2} \sum_i (D_i U)(D_i V)
\end{aligned}$$

Lemma 1.5.3. In every $T \in \tau^h$, the Laplacian operator can be expressed as :

$$\Delta = \frac{2}{3h^2} \sum_i D_{i,i} \tag{1.5.8}$$

Proof : It follows from equation (1.5.5) that

$$\Delta = \nabla\cdot\nabla = (\frac{2}{3h}\sum_i e_i D_i)\cdot(\frac{2}{3h}\sum_i e_i D_i)$$

$$= \frac{4}{9h^2}\sum_i (e_i\cdot e_i D_{i,i} + e_i\cdot e_{i+1} D_{i,i+1} + e_i\cdot e_{i-1} D_{i,i-1})$$

$$= \frac{4}{9h^2}\sum_i [D_{i,i} - \frac{1}{2}(D_{i,i+1} + D_{i,i-1})]$$

$$= \frac{4}{9h^2}\sum_i (D_{i,i} + \frac{1}{2}D_{i,i})$$

$$= \frac{2}{3h^2}\sum_i D_{i,i}$$

completing the proof.

1.6 SOBOLEV SPACE $H^k(\Omega)$

Denote by $H^k(\Omega)$, $k \geq 0$ the Sobolev space of real valued functions which together with their generalized derivatives up to the k^{th} order are square integrable over Ω [T1]. It is a linear subspace of $L^2(\Omega)$.

Denote by $(u,v) = \int_\Omega uv\, d\mu_\Omega = \sum_{T\in\tau} \mu_\Omega(T)\int_T uv\, d\mu_T$ the usual scalar product of the Hilbert space $L^2(\Omega)$.

Denote by $(u,v)_{k,T} = \sum_{|\alpha|\leq k} \frac{1}{h^{2|\alpha|}}\int_T (D^\alpha u)(D^\alpha v)\, d\mu_T$

by $(u,v)_{k,\Omega} = \sum_{T\in\tau} \mu_\Omega(T)(u,v)_{k,T}$, then the Sobolev space $H^k(\Omega)$ is a Hilbert space with the scalar product $(u,v)_{k,\Omega}$ [T1,p.55]. The corresponding Sobolev norm will be $\|u\|_{k,\Omega} = [(u,v)_{k,\Omega}]^{\frac{1}{2}}$.

Denote by $|u|_{k,T} = \{\sum_{|\alpha|=k} h^{-2k} \int_T (D^\alpha u)^2 \, d\mu_T\}^{\frac{1}{2}}$,

by $|u|_{k,\Omega} = \{\sum_{|\alpha|=k} h^{-2k} \int_\Omega (D^\alpha u)^2 \, d\mu_\Omega\}^{\frac{1}{2}}$ the Sobolev semi-norm of u on the triangle T and the domain Ω respectively.

1.7 PEANO-SARD KERNEL THEOREM AND ITS APPLICATION

Peano-Sard Kernel Theorem : Let Ω be a bounded polygonal domain. If $E : H^k(\Omega) \to H^k(\Omega)$ can be represented as $\int_\Omega \kappa f \, d\mu_\Omega$ for some $\kappa \in S^{n,\ell}(\Omega)$, and $E(f) = 0$ for all $f \in P^k(\Omega)$, then $\exists\ \kappa_\alpha \in S^{n+k,\ell+k}(\Omega)$, $|\alpha| = k$ s.t.

$$E(f) = \sum_{T\in\tau} \mu_\Omega(T) \int_T \left(\sum_{|\alpha|=k} \kappa_\alpha D^\alpha f \right) d\mu_T \tag{1.7.1}$$

A proof of the Theorem was given by P. Frederickson [F6].

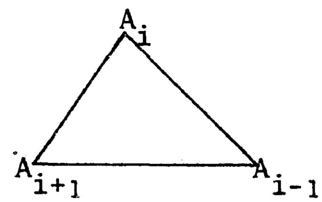
In the one dimensional case, we have the trapezoidal numerical quadrature for the integral $\int_a^b f\,dx$, which is exact for polynomials of degree ≤ 1, in the two dimensional case, we also have a similar numerical quadrature for the integral $\int_T f\,d\mu_T$ i.e. $\frac{1}{3}\sum_i f(A_i)$, where A_i, $i = 0, 1, 2$ are the three vertices of the triangle T. Clearly, this numerical quadrature is exact for polynomials of degree ≤ 1. By applying the Peano-Sard Kernel Theorem, we have the following lemma :

<u>Lemma 1.7.1.</u> If $f \in H^2(\Omega)$ and $E(f) = \int_T f \, d\mu_T - \frac{1}{3}\sum_i f(A_i)$,

then $E(f) = \sum_i \int_T \kappa_i D_{i,i} f \, d\mu_T$

$$\text{where } \kappa_i = -\frac{1}{12}(1-2\xi_i+2\xi_{i-1}\xi_{i+1}) \tag{1.7.2}$$

<u>Proof</u> : $\int_T f \, d\mu_T - \frac{1}{3}{}_i f(A_i) = \frac{1}{3}\sum_i \int_T [f-f(A_i)] \, d\mu_T$



$$= \frac{1}{3}\sum_i \int_T [f-f(A_i)] \; D_i[\tfrac{1}{2}(\xi_{i-1}-\xi_{i+1})] \, d\mu_T$$

$$= \frac{1}{3}\sum_i [2\int_{A_{i-1}}^{A_i} (f-f(A_i))\tfrac{1}{2}(\xi_{i-1}-\xi_{i+1}) \, dX -$$

$$2\int_{A_i}^{A_{i+1}} (f-f(A_i))\tfrac{1}{2}(\xi_{i-1}-\xi_{i+1}) \, dX - \int_T \tfrac{1}{2}(\xi_{i-1}-\xi_{i+1}) D_i f \, d\mu_T]$$

Since ξ_{i+1} and ξ_{i-1} vanish on $A_{i-1}A_i$ and A_iA_{i+1} respectively, we have

$$E(f) = \frac{1}{3}\sum_i [\int_{A_{i-1}}^{A_i} (f-f(A_i))(1-\xi_i) \, dX + \int_{A_i}^{A_{i+1}} (f-f(A_i))(1-\xi_i) \, dX +$$

$$\frac{1}{2}\int_T D_i(\xi_{i-1}\xi_{i+1}) D_i f \, d\mu_T]$$

$$= \frac{1}{3}\sum_i [\int_{A_{i-1}}^{A_i} (f-f(A_i)) D_{i+1}(\xi_i - \tfrac{1}{2}\xi_i^2) \, dX -$$

$$\int_{A_i}^{A_{i+1}} (f-f(A_i)) D_{i-1}(\xi_i - \tfrac{1}{2}\xi_i^2) \, dX + \int_{A_{i-1}}^{A_i} \xi_{i-1}\xi_{i+1} D_i f \, dX -$$

$$\int_{A_i}^{A_{i+1}} \xi_{i-1}\xi_{i+1} D_i f \, dX - \frac{1}{2}\int_T \xi_{i-1}\xi_{i+1} D_{i,i} f \, d\mu_T]$$

$$= \frac{1}{3}\sum_i [\,(f-f(A_i)(\xi_i - \frac{1}{2}\xi_i^2)\Big|_{A_{i-1}}^{A_i} - \int_{A_{i-1}}^{A_i} (\xi_i - \frac{\xi_i^2}{2})D_{i+1}f\ dX -$$

$$(f-f(A_i))(\xi_i - \frac{1}{2}\xi_i^2)\Big|_{A_i}^{A_{i+1}} + \int_{A_i}^{A_{i+1}} (\xi_i - \frac{1}{2}\xi_i^2)D_{i-1}f\ dX -$$

$$\frac{1}{2}\int_T \xi_{i-1}\xi_{i+1}D_{i,i}f\ d\mu_T]$$

$$= \frac{1}{3}\sum_i [-\int_{A_{i-1}}^{A_i} (\xi_i - \frac{1}{2}\xi_i^2)D_{i+1}f\ dX + \int_{A_i}^{A_{i+1}} (\xi_i - \frac{1}{2}\xi_i^2)D_{i-1}f\ dX -$$

$$\frac{1}{2}\int_T \xi_{i-1}\xi_{i+1}D_{i,i}f\ d\mu_T]$$

$$= \frac{1}{3}\sum_i [\int_{A_{i-1}}^{A_i} (\xi_i - \frac{1}{2}\xi_i^2)(D_i f + D_{i-1}f)\ dX -$$

$$\int_{A_i}^{A_{i+1}} (\xi_i - \frac{1}{2}\xi_i^2)(D_i f + D_{i+1}f)\ dX - \frac{1}{2}\int_T \xi_{i-1}\xi_{i+1}D_{i,i}f\ d\mu_T]$$

(1.7.3)

The line integrals in (1.7.3) can be rearranged into the following form :

$$E(f) = \frac{1}{3}\sum_i [\int_{A_{i-1}}^{A_i} (\xi_i - \frac{1}{2}\xi_i^2)D_i f\ dX + \int_{A_i}^{A_{i+1}} (\xi_{i+1} - \frac{1}{2}\xi_{i+1}^2)D_i f\ dX -$$

$$\int_{A_i}^{A_{i+1}} (\xi_i - \frac{1}{2}\xi_i^2)D_i f\ dX - \int_{A_{i-1}}^{A_i} (\xi_{i-1} - \frac{1}{2}\xi_{i-1}^2)D_i f\ dX -$$

$$\frac{1}{2}\int_T \xi_{i-1}\xi_{i+1}D_{i,i}f\ d\mu_T] \qquad (1.7.4)$$

By expressing $\xi_{i\pm 1}$ in the line integrals of (1.7.4) in terms of ξ_i and from Lemma 1.4.3., we obtain

$$E(f) = \frac{1}{3}\sum_i\Big[\frac{1}{2}\int_T(\xi_i - \frac{1}{2}\xi_i^2)D_{i,i}f\ d\mu_T - \frac{1}{2}\Big(\int_{A_{i-1}}^{A_i}(1-\xi_i^2)D_if\ dX -$$

$$\int_{A_i}^{A_{i+1}}(1-\xi_i^2)D_if\ dX\Big) - \frac{1}{2}\int_T\xi_{i-1}\xi_{i+1}D_{i,i}f\ d\mu_T\Big]$$

$$= \frac{1}{3}\sum_i\int_T\Big(\frac{1}{2}\xi_i - \frac{1}{4}\xi_i^2 - \frac{1}{4} + \frac{1}{4}\xi_i^2 - \frac{1}{2}\xi_{i-1}\xi_{i+1}\Big)D_{i,i}f\ d\mu_T$$

$$= \sum_i\int_T - \frac{1}{12}(1 - 2\xi_i + 2\xi_{i-1}\xi_{i+1})D_{i,i}f\ d\mu_T$$

completing the proof.

Unlike the one dimensional case, the Peano-Sard kernels for the linear interpolation error functional are not unique. We shall derive two different forms of Peano-Sard kernels for the linear spline interpolation remainder. These kernels will be applied to the finite element error analysis in Chapter 3.

Let $f : \Omega \to R$, then the piecewise linear interpolation of f on each triangular element $T = A_1A_2A_3$ is given by

$$f_I(A_0) = \sum_i x_i f(A_i)$$

where (x_1,x_2,x_3) is the Barycentric Coordinates of a point A_0 w.r.t. T.

In order to obtain the kernels of the error functional $E(f,A_0) = f(A_0) - \sum_i x_i f(A_i)$, we need the relationships between the Barycentric Coordinates of a point P w.r.t. the triangles T and $T_i = A_0A_{i+1}A_{i-1}$.

Denote by (ξ_1,ξ_2,ξ_3) the Barycentric Coordinates of P w.r.t. T, by $(\xi_i^{T_i},\xi_{i+1}^{T_i},\xi_{i-1}^{T_i})$ the Barycentric Coordinates of P w.r.t. T_i. Then it follows from (1.3.4) that

$$\begin{cases} \xi_i = \xi_i^{T_i} x_i \\ \xi_{i+1} = \xi_i^{T_i} x_{i+1} + \xi_{i+1}^{T_i} \\ \xi_{i-1} = \xi_i^{T_i} x_{i-1} + \xi_{i-1}^{T_i} \end{cases} \qquad (1.7.5)$$

Denote by $F(\xi_1,\xi_2,\xi_3)$ an expression of f w.r.t. the triangle T, by $F^{T_i}(\xi_i^{T_i},\xi_{i+1}^{T_i},\xi_{i-1}^{T_i})$ an expression of f w.r.t. the triangle $T_i = A_0A_{i+1}A_{i-1}$.

As shown in Fig. 1.7.1, let $D_{e_i} f$ be the normalized derivative of f in the direction $\dot{A_i A_0}$, then from (1.4.1) we have

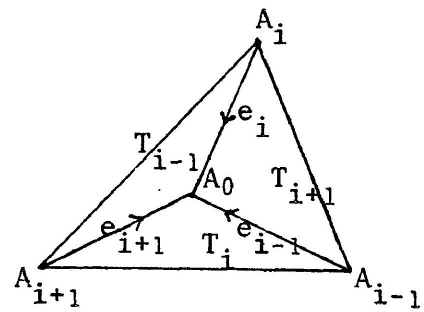


Fig. 1.7.1

$$D_{e_{i+1}} f = \frac{\partial F^{T_i}}{\partial \xi_i^{T_i}} - \frac{\partial F^{T_i}}{\partial \xi_{i+1}^{T_i}}$$

$$= \sum_j \frac{\partial F}{\partial \xi_j} \frac{\partial \xi_j}{\partial \xi_i^{T_i}} - \sum_j \frac{\partial F}{\partial \xi_j} \frac{\partial \xi_j}{\partial \xi_{i+1}^{T_i}}$$

$$= \sum_j \frac{\partial F}{\partial \xi_j} D_{e_{i+1}} \xi_j \tag{1.7.6}$$

By substituting (1.7.5) into (1.7.6), we have

$$D_{e_{i+1}} f = \frac{\partial F}{\partial \xi_i} x_i + \frac{\partial F}{\partial \xi_{i+1}} (x_{i+1} - 1) + \frac{\partial F}{\partial \xi_{i-1}} x_{i-1}$$

$$= \frac{\partial F}{\partial \xi_i} x_i - \frac{\partial F}{\partial \xi_{i+1}} (x_i + x_{i-1}) + \frac{\partial F}{\partial \xi_{i-1}} x_{i-1}$$

$$= x_i \left(\frac{\partial F}{\partial \xi_i} - \frac{\partial F}{\partial \xi_{i+1}}\right) + x_{i-1} \left(\frac{\partial F}{\partial \xi_{i-1}} - \frac{\partial F}{\partial \xi_{i+1}}\right)$$

$$= \begin{cases} -x_i D_{i-1} f + x_{i-1} D_i f & \text{or} \\ x_i (D_i f + D_{i+1} f) + x_{i-1} D_i f & \text{or} \\ -x_i D_{i-1} f - x_{i-1} (D_{i-1} f + D_{i+1} f) \end{cases} \tag{1.7.7}$$

We observe that, though the representations of F and F^{T_i} are not unique, the final forms of $D_{e_{i+1}} f$ are independent of F and F^{T_i}.

Now we have three different expressions to resolve $D_{e_i} f$ in terms of the derivatives $D_i f$ and $D_{i\pm 1} f$ i.e.

$$D_{e_i} f \equiv \begin{cases} x_{i+1} D_{i-1} f - x_{i-1} D_{i+1} f & (1.7.8) \\ (x_{i-1} + x_{i+1}) D_{i-1} f + x_{i-1} D_i f & (1.7.9) \\ -(x_{i-1} + x_{i+1}) D_{i+1} f - x_{i+1} D_i f & (1.7.10) \end{cases}$$

If $f \in H^2(\Omega)$, then the linear spline $f_I(A_0) = \sum_i x_i f(A_i)$ interpolates f in the triangle T, and the error functional $E(f,A_0) = f(A_0) - \sum_i x_i f(A_i)$ is exact for polynomials of degree ≤ 1, by Peano-Sard Kernel Theorem, there exist kernels κ_α, s.t.

$$E(f,A_0) = \int_T \sum_{|\alpha|=2} \kappa_\alpha D^\alpha f \, d\mu_T$$

The problem is how to construct the kernels κ_α.

<u>Claim</u> : The kernels κ_α are piecewise constant (functions of x_i, i=1, 2, 3 only) and it can be written in the form :

$$E(f,A_0) = \sum_i \int_{T_i} (\kappa_i^{i-1} D_{i,i-1} f + \kappa_i^{i+1} D_{i,i+1} f) \, d\mu_{T_i}$$

$$= \sum_i \{ 2\kappa_i^{i-1} [\int_{A_{i-1}}^{A_0} D_{i-1} f \, dX - \int_{A_{i+1}}^{A_0} D_{i-1} f \, dX] +$$

$$2\kappa_i^{i+1} [\int_{A_{i-1}}^{A_0} D_{i+1} f \, dX - \int_{A_{i+1}}^{A_0} D_{i+1} f \, dX] \}$$

(1.7.11)

Rearranging the sums in the equation (1.7.11), we have

$$E(f,A_0) = 2\sum_i \int_{A_i}^{A_0} (\kappa_{i+1}^{i} D_i f - \kappa_{i-1}^{i+1} D_{i+1} f + \kappa_{i+1}^{i-1} D_{i-1} f - \kappa_{i-1}^{i} D_i f)\, dX \tag{1.7.12}$$

From (1.7.9) and (1.7.10), we have

$$D_{i-1} f = \frac{D_{e_i} f - x_{i-1} D_i f}{x_{i-1} + x_{i+1}}$$

$$D_{i+1} f = - \frac{D_{e_i} f + x_{i+1} D_i f}{x_{i-1} + x_{i+1}}$$

Substituting these into (1.7.12), we obtain

$$E(f,A_0) = 2\sum_i \int_{A_i}^{A_0} [(\kappa_{i+1}^{i} - \kappa_{i-1}^{i}) D_i f + \frac{\kappa_{i-1}^{i+1}(D_{e_i} f + x_{i+1} D_i f)}{x_{i-1} + x_{i+1}} +$$

$$\frac{\kappa_{i+1}^{i-1}(D_{e_i} f - x_{i-1} D_i f)}{x_{i-1} + x_{i+1}}]\, dX$$

$$= 2\sum_i \int_{A_i}^{A_0} [\frac{(\kappa_{i-1}^{i+1} + \kappa_{i+1}^{i-1}) D_{e_i} f}{x_{i-1} + x_{i+1}} +$$

$$(\kappa_{i+1}^{i} - \kappa_{i-1}^{i} + \frac{x_{i+1}\kappa_{i-1}^{i+1} - x_{i-1}\kappa_{i+1}^{i-1}}{x_{i-1} + x_{i+1}}) D_i f]\, dX$$

We want

$$\begin{cases} \dfrac{\kappa_{i+1}^{i-1} + \kappa_{i-1}^{i+1}}{x_{i-1} + x_{i+1}} = \dfrac{1}{2}x_i \\ \kappa_{i+1}^{i} - \kappa_{i-1}^{i} + \dfrac{x_{i+1}\kappa_{i-1}^{i+1} - x_{i-1}\kappa_{i+1}^{i-1}}{x_{i-1} + x_{i+1}} = 0 \end{cases}$$

It follows that

$$\begin{cases} \kappa_{i+1}^{i-1} + \kappa_{i-1}^{i+1} = \frac{1}{2}x_i(x_{i-1} + x_{i+1}) \\ (x_{i-1} + x_{i+1})(\kappa_{i+1}^{i} - \kappa_{i-1}^{i}) = x_{i-1}\kappa_{i+1}^{i-1} - x_{i+1}\kappa_{i-1}^{i+1} \end{cases}$$

By solving the above system of linear equation for $i = 1, 2, 3,$ we obtain

$$\kappa_i^{i+1} = \frac{1}{2}x_i x_{i-1} \tag{1.7.13}$$

Thus, we have

$$E(f,A_0) = \sum_i \int_{T_i} (\frac{1}{2}x_i x_{i+1} D_{i,i-1}f + \frac{1}{2}x_i x_{i-1}\ D_{i,i+1}f)\ d\mu_{T_i} \tag{1.7.14}$$

We shall derive the kernel of the same error functional $E(f,A_0)$ by a different approach and obtain another different kernel of $E(f,A_0)$.

$$E(f,A_0) = f(X) - \sum_i x_i f(A_i) \qquad (1.7.15)$$

$$= \sum_i x_i (f(X)-f(A_i))$$

$$= \sum_i x_i \int_{A_i}^{A_0} D_{e_i} f(\S)\, d\S$$

It follows from (1.7.8) that

$$E(f,A_0) = \sum_i x_i \int_{A_i}^{A_0} (x_{i+1} D_{i-1} f - x_{i-1} D_{i+1} f)\, dX$$

$$= \sum_i (x_i x_{i+1} \int_{A_i}^{A_0} D_{i-1} f\, dX - x_i x_{i-1} \int_{A_i}^{A_0} D_{i+1} f\, dX)$$

Rearranging the sums of the above line integrals, we have

$$E(f,A_0) = \sum_i x_{i-1} x_{i+1} (\int_{A_{i+1}}^{A_0} D_i f\, dX - \int_{A_{i-1}}^{A_0} D_i f\, dX)$$

$$E(f,A_0) = \sum_i (- \frac{x_{i-1} x_{i+1}}{2}) \int_{T_i} D_{i,i} f\, d\mu_{T_i} \qquad (1.7.16)$$

We observe that (1.7.16) can be written in the form

$$E(f,A_0) = \sum_i \int_{T_i} (\frac{1}{2} x_{i-1} x_{i+1} D_{i,i-1} f + \frac{1}{2} x_{i-1} x_{i+1} D_{i,i+1} f)\, d\mu_{T_i}$$

$$(1.7.17)$$

The kernels of f in (1.7.14) and (1.7.17) are not the same, so this example shows that the kernels are not unique.

By equating the equations (1.7.14) and (1.7.17), we obtain the following identity :

If $f \in H^2(\Omega)$, then

$$\sum_i \int_{T_i} [x_{i+1}(x_i - x_{i-1}) D_{i,i-1} f + x_{i-1}(x_i - x_{i+1}) D_{i,i+1} f] \, d\mu_{T_i} = 0$$

CHAPTER 2

FINITE ELEMENT SOLUTION TO THE SECOND ORDER ELLIPTIC PROBLEMS

2.1 INTRODUCTION

Consider the second order elliptic boundary value problem ([S4],[A1],[B4]), defined in a bounded open domain Ω with polygonal boundary $\partial\Omega$ by

$$\begin{cases} Lu = -\nabla\cdot(p\nabla u) + qu = f & \text{in } \Omega & (2.1.1) \\ u = g & \text{on } \partial\Omega & (2.1.2) \end{cases}$$

This differential equation arises in a variety of physical contexts, for example, the equation (2.1.1) is satisfied by the transverse deflection $u(X)$ of a membrane under uniform lateral tension T, which supports a load of $Tf(X)$ per unit area.

Under the assumptions p, q are smooth functions and

$$\begin{cases} p \geq p_{min} > 0 \\ q \geq 0 \end{cases} \quad \text{in } \Omega, \qquad (2.1.3)$$

the differential operator $L = -\nabla\cdot p\nabla + q$ is a 1-1 continuous linear operator ([S3],[T1]) mapping $H_g^2(\Omega)$ onto $H^0(\Omega)$, where $H_g^2(\Omega)$ is the solution space defined by

$$H_g^2(\Omega) = \{u \in H^2(\Omega) : u=g \text{ on } \partial\Omega\} .$$

In general, if $g \neq 0$, then $H^2_g(\Omega)$ is not a linear space, but an affine subspace of $H^2(\Omega)$.

In particular, if $p = 1$ and $q = 0$, the equation (2.1.1) reduces to the Poisson equation

$$-\Delta u = f \tag{2.1.4}$$

2.2 THE VARIATIONAL FORM OF THE PROBLEM

The problem of solving a boundary value problem often turns out to be equivalent to the problem of minimizing a certain quadratic functional ([B4],[A1]) .

The quadratic functional related to the linear equation (2.1.1) is given by

$$I(v) = \int_\Omega (p\nabla v\cdot\nabla v + qv^2 - 2fv)\, d\mu_\Omega \tag{2.1.5}$$

The solution of the differential problem $Lu = f$ is expected to coincide with the function u that minimizes I. Since the integral (2.1.5) involves no second derivatives, the class of functions over which the integral $I(v)$ is to be minimized is enlarged to the space of admissible functions defined by

$$H^1_g(\Omega) = \{u \in H^1(\Omega) \;:\; u = g \quad \text{on} \quad \partial\Omega\}$$

We observe that the admissible space $H^1_g(\Omega)$ is an affine space, and can be written as $H_0^1(\Omega) + g$, where $H_0^1(\Omega)$ denotes the linear suspace $\{u \in H^1(\Omega) : u = 0 \quad \text{on } \partial\Omega\}$

Before we proceed further, the first step is to check that a solution u to the differential problem does minimize I(v).

Let u be an admissible function of the integral I(v), and v be any function in $H_0^1(\Omega)$. For every ε in R, the function $u+\varepsilon v$ is still an admissible function of I(v) and we have

$$I(u+\varepsilon v) = \int_\Omega [p\nabla(u+\varepsilon v)\cdot\nabla(u+\varepsilon v)+q(u+\varepsilon v)^2-2f(u+\varepsilon v)]\ d\mu_\Omega$$

$$= \int_\Omega (p\nabla u\cdot\nabla u+qu^2-2fu)\ d\mu_\Omega+2\varepsilon\int_\Omega (p\nabla u\cdot\nabla v+quv-fv)\ d\mu_\Omega+$$

$$\varepsilon^2\int_\Omega (p\nabla v\cdot\nabla v+qv^2)\ d\mu_\Omega$$

It follows that

$$\frac{dI(u+\varepsilon v)}{d\varepsilon} = 2\int_\Omega (p\nabla u\cdot\nabla v+quv-fv)\ d\mu_\Omega+2\varepsilon\int_\Omega (p\nabla v\cdot\nabla v+qv^2)\ d\mu_\Omega$$

and

$$\frac{d^2I(u+\varepsilon v)}{d\varepsilon^2} = 2\int_\Omega (p\nabla v\cdot\nabla v+qv^2)\ d\mu_\Omega$$

Since $p > 0$, $q \geq 0$ and $\nabla v \cdot \nabla v \geq 0$ for all $v \in H_0^1(\Omega)$ we have

$$\frac{d^2 I(u+\varepsilon v)}{d\varepsilon^2} \geq 0 \qquad \forall \ v \ \text{in} \ H_0^1(\Omega)$$

Thus, an admissible function u minimizes $I(v)$ iff the first variation

$$\left.\frac{dI}{d\varepsilon}(u+\varepsilon v)\right|_{\varepsilon=0}$$

vanishes for all v in $H_0^1(\Omega)$, that is, if and only if

$$\int_\Omega (p\nabla u \cdot \nabla v + quv - fv)\ d\mu_\Omega = 0 \tag{2.1.6}$$

By Green's Theorem ([W1,p.346],[H2]) equation (2.1.6) is equivalent to

$$\int_\Omega [-\nabla \cdot (p\nabla u) + qu - f]v\ d\mu_\Omega - \int_{\partial\Omega} \frac{\partial u}{\partial n} v\ ds = 0 \tag{2.1.7}$$

where $\frac{\partial u}{\partial n}$ is the outward normal derivative of u on $\partial\Omega$.

Since $v \in H_0^1(\Omega)$, the line integral of (2.1.7) vanishes and we have

$$\int_\Omega [-\nabla \cdot (p\nabla u) + qu - f]v\ d\mu_\Omega = 0 \tag{2.1.8}$$

This holds for all $v \in \overset{1}{H}_0(\Omega)$ iff

$$-\nabla\cdot(p\nabla u) + qu = f$$

Thus, the elliptic equation (2.1.1) turns out to be the Euler equation for the problem of minimizing the integral $I(v)$. Also, the second variation $\left.\frac{d^2 I(u+\varepsilon v)}{d\varepsilon^2}\right|_{\varepsilon=0}$ is positive unless v is constant, which implies by the boundary condition, that v vanishes identically. Thus u will be the unique function which minimizes the quadratic function (2.1.5).

2.3 ENERGY INNER PRODUCT

Define a bilinear expression on $\overset{1}{H}_0(\Omega)\times\overset{1}{H}_0(\Omega)$ by

$$a(u,v) = \int_\Omega (p\nabla u\cdot\nabla v+quv)\, d\mu_\Omega \qquad (2.3.1)$$

It is easy to check that $a(\cdot,\cdot)$ has the following properties :

(i) $a(u_1+u_2,v) = a(u_1,v) + a(u_2,v)$

(ii) $a(u,v) = a(v,u)$

(iii) $a(\lambda u,v) = \lambda a(u,v)$ for all $\lambda \in R$

(iv) $a(u,u) \geq 0$ for all $u \in \overset{1}{H}_0(\Omega)$

(v) $a(u,u) = 0$ iff $u = 0$

Thus $a(\cdot,\cdot)$ is an inner product on the space $\overset{1}{H}_0(\Omega)\times\overset{1}{H}_0(\Omega)$. This inner product is referred to as the energy inner product, and the norm defined by $\|u\|_a = [a(u,u)]^{\frac{1}{2}}$ will be referred to as the energy norm. In particular, if $p = 1$, $q = 0$, the corresponding energy norm will be denoted by $\|u\|_\Delta$.

Theorem 2.3.1. The energy norm $\|u\|_a$ is equivalent to the Sobolev norm $\|u\|_{1,\Omega}$.

The Theorem is proved by the following two lemmas.

Lemma 2.3.1. There is a constant $\rho > 0$. Such that

$$\|u\|_a \le \rho\|u\|_{1,\Omega}$$

Proof : From Lemma 1.5.2. we have

$$\nabla u\cdot\nabla u = \frac{2}{3h^2}\sum_i (D_i u)^2$$

It follows that

$$\begin{aligned} a(u,u) &= \sum_{T\in\tau^h} \mu_\Omega(T)\int_T [p(\frac{2}{3h^2})\sum_i (D_i u)^2 + qu^2]\, d\mu_T \\ &\le \sum_{T\in\tau^h} \mu_\Omega(T)\max_{X\in\Omega}(\frac{2p(X)}{3}, q(X))\int_T (\frac{1}{h^2}\sum_i (D_i u)^2 + u^2)\, d\mu_T \\ &= \max_{X\in\Omega}(\frac{2p(X)}{3}, q(X))\, \|u\|^2_{1,\Omega} \end{aligned}$$

Since $p \geq p_{min} > 0$ and $q \geq 0$ in Ω, we have

$$\max\left(\frac{2p(X)}{3}, q(X)\right) > 0$$

Letting $\rho = [\max_{X\in\Omega}(\frac{2p(X)}{3}, q(X))]^{\frac{1}{2}}$, we get

$$\|u\|_a = [a(u,u)]^{\frac{1}{2}} \leq \rho \|u\|_{1,\Omega} \quad ,$$

completing the proof.

Lemma 2.3.2. There is a constant $\sigma > 0$. Such that

$$\sigma \|u\|_{1,\Omega} \leq \|u\|_a$$

Proof : $a(u,u) = \sum_{T\in\tau^h} \mu_\Omega(T) \int_T [p(\frac{2}{3h^2}) \sum_i (D_i u)^2 + qu^2] \, d\mu_T$

$$\geq \min_{X\in\Omega}\left(\frac{2p(X)}{3}\right) \sum_{T\in\tau^h} \mu_\Omega(T) \int_T \frac{1}{h^2} \sum_i (D_i u)^2 \, d\mu_T$$

By the Poincaré inequality ([S4], [P1]), there is a constant $\tilde{\sigma} > 0$, such that

$$\int_\Omega \nabla u \cdot \nabla u \, d\mu_\Omega \geq \tilde{\sigma} \int_\Omega u^2 \, d\mu_\Omega \qquad \text{for all} \quad u \in H_0^1(\Omega) \ .$$

Since $p \geq p_{min} > 0$ in Ω, we have

$$\sigma = [\frac{1}{2} \min(\tilde{\sigma}, \min_{X\in\Omega}(\frac{2p(X)}{3}))]^{\frac{1}{2}} > 0$$

It follows that

$$||u||_a = [a(u,u)]^{\frac{1}{2}}$$

$$\geq \sigma[\sum_{T\in\tau^h} \mu_\Omega(T)\int_T (\frac{1}{h^2}\sum_i (D_i u)^2+u^2)\ d\mu_T]^{\frac{1}{2}}$$

$$= \sigma||u||_{1,\Omega}$$

completing the proof.

2.4 THE RITZ-GALERKIN METHOD

Consider the equation

$$Lu = f \qquad (2.4.1)$$

Assume (2.4.1) has a solution in the Hilbert space H with the inner product $(\cdot,\cdot)$. If L is linear, symmetric and positive definite. Then as we have discussed in the last section, solving of (2.4.1) is equivalent to minimization of the quadratic functional

$$I(v) = (Lv,v) - 2(f,v) \qquad (2.4.2)$$

over an admissible space H_B.

The Ritz method ([S3], [P1], [B6], [A1]) is to replace H_B by a finite dimensional subspace S^h contained in H_B. The elements v^h of S^h are called trial functions. If ϕ_i, $i = 1,\cdots,n$ are the n basis elements of S^h, then every member of S^h can be written as

$$v^h = \sum_{i=1}^{n} \lambda_i \phi_i \tag{2.4.3}$$

By substituting (2.4.3) into $I(v^h)$ and letting the derivatives $\frac{\partial I}{\partial \lambda_i}$ be zero for $i = 1, 2, \cdots, n$. The Ritz method turns out to be the solution of a system of linear equations of the form

$$\sum_{j=1}^{n} \lambda_j (L\phi_j, \phi_i) = (f, \phi_i) \quad \text{for } i=1, 2, \cdots, n \tag{2.4.4}$$

Since the linear operator L is symmetric and positive definite, the solution of (2.4.4) exists and is unique.

The main weakness of the Ritz method is the fact that it is applicable only to equations with symmetric and positive definite linear operators. Another method, called the Galerkin method is free from this constraint. We shall describe this method with an example of solving the equation (2.4.1).

An element $u \in H$ is called a weak (or 'generalized') solution of the problem (2.4.1) if

$$(Lu, v) = (f, v) \quad \text{for all } v \in H$$

The Galerkin approximation to the problem $Lu = f$ is to seek a weak solution in a finite dimensional subspace S^h of H ([S4],[P1],[B4],[M1]). Thus, if ϕ_i, $i=1,\cdots,n$ are the n basis elements of S^h, it is sufficient to find $u^h \in S^h$, such that

$$(Lu^h,\phi_i) = (f,\phi_i) \quad \text{for all} \quad i=1, 2,\cdots, n \,. \qquad (2.4.5)$$

It is easy to check that for a linear, symmetric and positive definite operator L, the two systems of equations (2.4.4) and (2.4.5) are identical. Thus the Galerkin method is a generalization of the Ritz method.

The linear operator $L = -\nabla\cdot p\nabla + q$ defined in (2.1.1) is linear and symmetric. As we have proved in Section 2.2, the inner product (Lu,v) is the same as the energy inner product $a(u,v)$, and from the result of Lemma 2.3.2. we know that L is positive definite. Thus, for this linear operator L, the Ritz method and the Galerkin method are equivalent, we shall refer to it as the Ritz-Galerkin method.

Denote by S^h a finite dimensional subspace of $H^1(\Omega)$, and by $\{\phi_i\}_{i=1}^n$ the n basis elements of S^h.

The Ritz-Galerkin solution to the problem (2.1.1) thus requires only the solution of the system of linear equations :

$$\int_\Omega (p\nabla u^h\cdot\nabla\phi_i+qu^h\phi_i)\ d\mu_\Omega = \int_\Omega f\phi_i\ d\mu_\Omega \qquad (2.4.6)$$

for $i=1, 2,\cdots, n$,

where $$u^h = g + \sum_{i=1}^n \lambda_i\phi_i \qquad (2.4.7)$$

2.5 RITZ-GALERKIN METHOD WITH TRIANGULAR LINEAR ELEMENTS

Given an equilateral triangulation τ^h of Ω, the simplest and most basic of all trial functions is the triangular linear elements. The trial function is linear inside each triangle and continuous across each edge ([S4],[P1],[C3]). Denote by $S_g^{1,0}$ the affine subspace define by

$$S_g^{1,0} = \{\phi \in S^{1,0} : \phi = g \quad \text{on } \partial\Omega\}.$$

For every element X_α of $\mathring{\Omega}_h$, let ϕ_α be the trial function which equals 1 at X_α and zero at all other nodes. Then these pyramid functions ϕ_α form a basis for the trial space $S_g^{1,0}$. The dimension of $S_g^{1,0}$ equals to the number of elements in $\mathring{\Omega}_h$.

Denote by $(\xi_\alpha, \xi_\beta, \xi_\gamma)$ the Barycentric Coordinates of a point X w.r.t. the triangle $T = X_\alpha X_\beta X_\gamma$.

The basis function $\phi_\alpha(X)$ can be expressed as

$$\phi_\alpha(X) = \begin{cases} \xi_\alpha & \text{if } X \in X_\alpha + H \\ 0 & \text{otherwise} \end{cases}$$

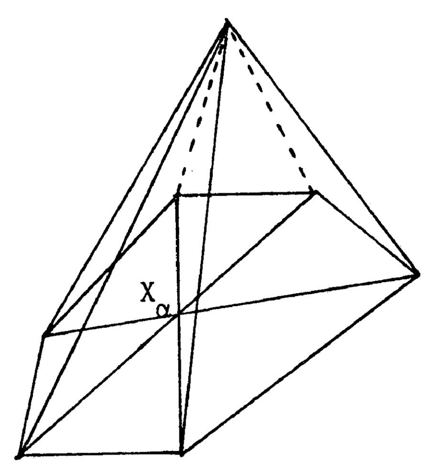
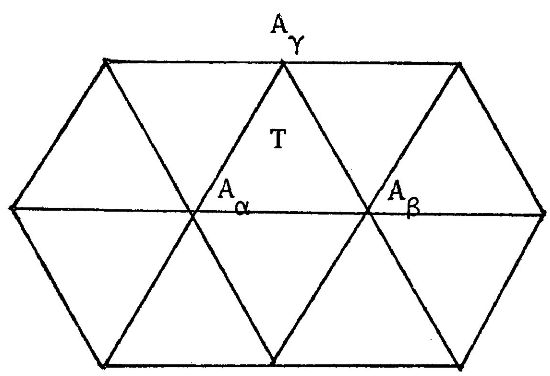


Fig. 2.5.1

To construct the Ritz-Galerkin approximation with triangular elements, we need the following Lemma.

<u>Lemma 2.5.1.</u> $$\int_\Omega \nabla\phi_\alpha \cdot \nabla\phi_\beta \, d\mu_\Omega = \frac{4\mu_\Omega(T)}{3h^2} \begin{cases} -1 & \text{if } |\alpha-\beta|=1 \\ 6 & \text{if } \alpha=\beta \\ 0 & \text{otherwise} \end{cases} \tag{2.5.1}$$

Proof :



As shown in the above figure, let $T=A_\alpha A_\beta A_\gamma$ be a triangle of the hexagon $A_\alpha + H$, then we have

$$\phi_\alpha = \xi_\alpha$$

and

$$D_\sigma\phi_\beta \begin{cases} 1 & \text{if } (\sigma,\beta) \in \{(\alpha,\gamma),(\gamma,\beta),(\beta,\alpha)\} \\ -1 & \text{if } (\sigma,\beta) \in \{(\alpha,\beta),(\beta,\gamma),(\gamma,\alpha)\} \\ 0 & \text{otherwise} \end{cases}$$

From Lemma 1.5.2., we get

$$\int_{\Omega} \nabla\phi_\alpha \cdot \nabla\phi_\beta \, d\mu_T = \sum_{T\in\tau^h} \mu_\Omega(T) \int_T \frac{2}{3h^2} \sum_\sigma (D_\sigma\phi_\alpha)(D_\sigma\phi_\beta) \, d\mu_T$$

$$= \begin{cases} 2\mu_\Omega(T) \int_T \frac{2}{3h^2}(-1) \, d\mu_T & \text{if } |\alpha-\beta|=1 \\ 6\mu_\Omega(T) \int_T \frac{2}{3h^2}(2) \, d\mu_T & \text{if } \alpha=\beta \\ 0 & \text{otherwise} \end{cases}$$

$$= \frac{4\mu_\Omega(T)}{3h^2} \begin{cases} -1 & \text{if } |\alpha-\beta|=1 \\ 6 & \text{if } \alpha=\beta \\ 0 & \text{otherwise} \end{cases}$$

completing the proof.

To construct the Ritz-Galerkin solution of (2.1.1) with the boundary condition $u = g$ on $\partial\Omega$, it is convenient to express the minimizing function u^h in terms of ϕ_α as

$$u^h = \sum_{\alpha\in\Gamma_h} \lambda_\alpha \phi_\alpha \tag{2.5.2}$$

We observe that only those interior parameters λ_α in the equation (2.5.2) are to be determined. For those nodes which lie on the boundary $\partial\Omega$,

$$\lambda_\alpha = g(X_\alpha) \ .$$

By substituting the equation (2.5.2) into (2.4.6), we have

$$\sum_{\beta\in\Gamma_h} \lambda_\beta \int_\Omega (p\nabla\phi_\alpha \cdot \nabla\phi_\beta + q\phi_\alpha\phi_\beta)\, d\mu_\Omega = \int_\Omega f\phi_\alpha \, d\mu_\Omega \qquad (2.5.3)$$

for $\alpha \in \mathring{\Gamma}_h$

It follows from Lemma 2.5.1 that the system of linear equations (2.5.3) becomes

$$\sum_{\beta\in\Gamma_h} \lambda_\beta L_{\alpha,\beta} = \int_\Omega f\phi_\alpha \, d\mu_\Omega \quad \text{for} \quad \alpha \in \mathring{\Gamma}_h \qquad (2.5.4)$$

where

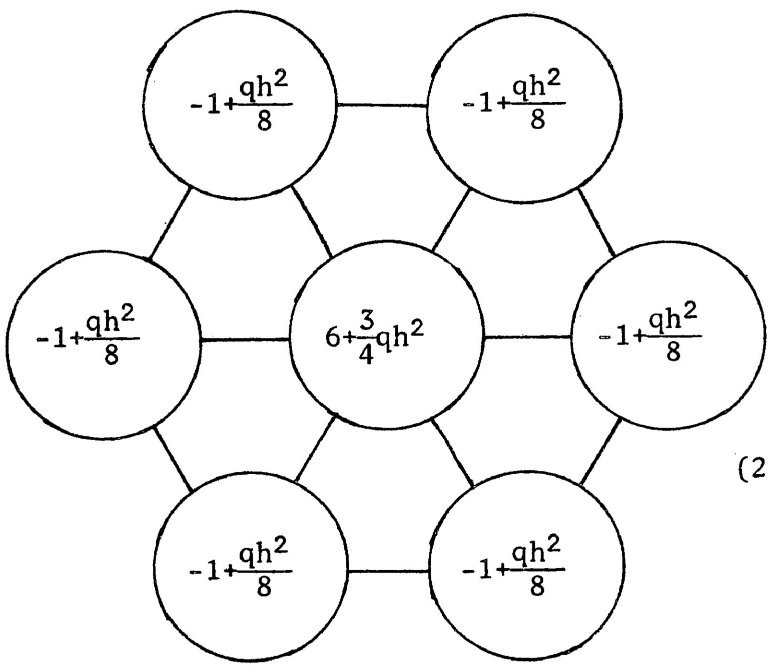
$$L_{\alpha,\beta} = \begin{cases} \displaystyle\int_{T_\alpha \cup T_\beta} [-\frac{2}{3h^2}p(X) + \phi_\alpha\phi_\beta q(X)]d\mu_\Omega & \text{if } |\alpha-\beta|=1 \\ \text{where } T_\alpha \text{ and } T_\beta \text{ are the two triangular elements in } \tau^h \\ \text{having the common side } X_\alpha X_\beta \\ \displaystyle\int_{X_\alpha + H} [\frac{4}{3h^2}p(X) + \phi_\alpha^2 q(X)]\, d\mu_\Omega & \text{if } \alpha=\beta \\ 0 \end{cases}$$

In particular, if $p=1$ and q is a constant, then

$$
L_{\alpha,\beta} = \begin{cases} \left(-\frac{4p}{3h^2} + \frac{q}{6}\right)\mu_\Omega(T) & \text{if } |\alpha-\beta|=1 \\ \left(\frac{8p}{h^2} + q\right)\mu_\Omega(T) & \text{if } \alpha=\beta \\ 0 & \text{otherwise} \end{cases}
$$

Expressed diagrammatically, the discrete linear operator L^h associated with the continuous operator $L = -\Delta + q$ has a representation of the form :

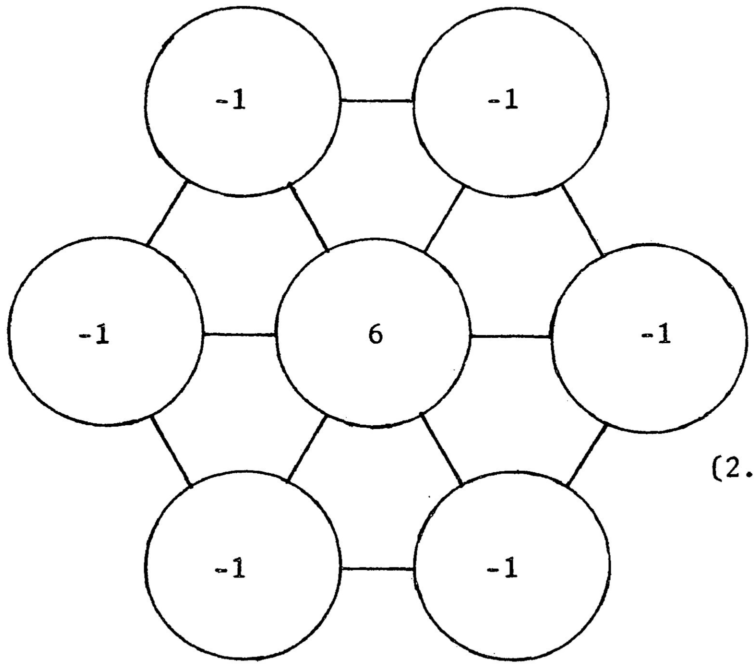
$$\frac{3h^2}{4\mu_\Omega(T)}L^h \ :$$



(2.5.5)

If $q = 0$, L becomes the Laplacian operator $-\Delta$. The associated discrete Laplacian operator L^h has a representation of the form :

$$\frac{3h^2}{4\mu_\Omega(T)}L^h :$$



(2.5.6)

2.6 NUMERICAL QUADRATURE FORMULAS

For arbitary p, q and f, the integrals in the expression (2.5.3) cannot be computed exactly, and some numerical quadrature will be necessary to approximate these integrals.

In this section, we shall derive some numerical quadrature formulas for the following four types of integrals :

(i) $F_\alpha = \int_\Omega f\phi_\alpha \, d\mu_\Omega$

(ii) $Q_\alpha = \int_\Omega q\phi_\alpha^2 \, d\mu_\Omega$

(iii) $P_{\alpha,\beta} = \int_{T_\alpha \cup T_\beta} p \, d\mu_\Omega$

(iv) $Q_{\alpha,\beta} = \int_\Omega q\phi_\alpha\phi_\beta \, d\mu_\Omega$

The corresponding numerical quadrature will be denoted by $\tilde{F}_\alpha$, $\tilde{Q}_\alpha$, $\tilde{P}_{\alpha,\beta}$ and $\tilde{Q}_{\alpha,\beta}$ respectively.

We observe that the integrals F_α and Q_α have support over the hexagon $X_\alpha + H$. The simplest numerical quadrature is the 1-point formula, that is, F_α and Q_α are approximated by $af(X_\alpha)$ and $bq(X_\alpha)$ respectively, where a and b are two constants to be determined.

To obtain the values of a and b, we may require that they be exact for constants f and q, that is

$$\begin{cases} \int_\Omega \phi_\alpha \, d\mu_\Omega - a = 0 \\ \int_\Omega \phi_\alpha^2 \, d\mu_\Omega - b = 0 \end{cases}$$

It follows that

$$a = \int_\Omega \phi_\alpha \, d\mu_\Omega = 6\mu_\Omega(T)\int_T \xi_\alpha \, d\mu_T = 2\mu_\Omega(T)$$

and

$$b = \int_\Omega \phi_\alpha^2 \, d\mu_\Omega = 6\mu_\Omega(T)\int_T \xi_\alpha^2 \, d\mu_T = \mu_\Omega(T)$$

By the symmetric form of the integrals $\int_\Omega f\phi_\alpha \, d\mu_\Omega$ and

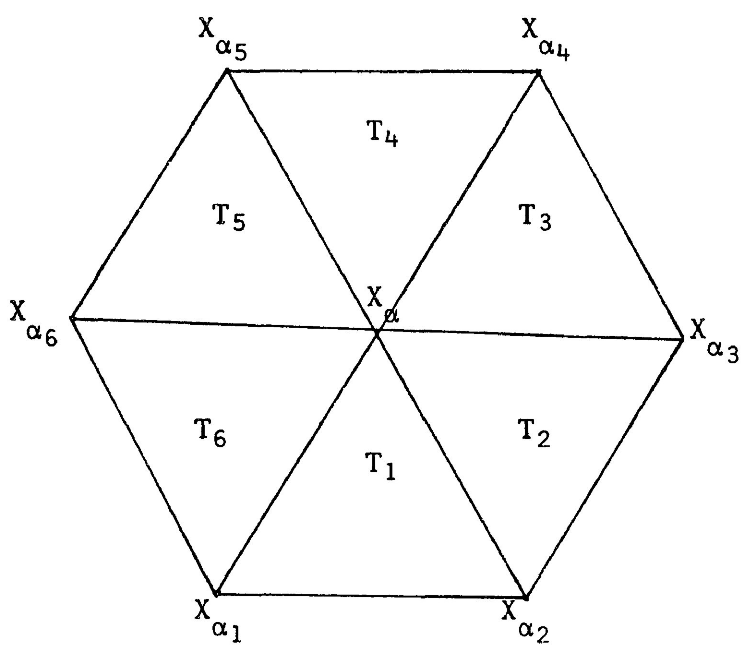
$\int_\Omega q\phi_\alpha^2 \, d\mu_\Omega$, it is easy to verify that the two numerical quadratures $\tilde{F}_\alpha = 2\mu_\Omega(T)f(X_\alpha)$ and $\tilde{Q}_\alpha = \mu_\Omega(T)q(X_\alpha)$ are exact for all polynomials of degree 1.

To obtain numerical quardrature with higher order of accuracy, we require the following Lemma :

<u>Lemma 2.6.1.</u> Let ψ be a quadratic polynomial which takes the value 1 along the edges $X_{\alpha_6}X_{\alpha_1}$ and $X_{\alpha_3}X_{\alpha_4}$ and vanishes along the line $X_{\alpha_2}X_{\alpha_5}$ of the hexagon $X_\alpha + H$. Then

(i) $\displaystyle\int_\Omega \psi\phi_\alpha \, d\mu_\Omega = \frac{\mu_\Omega(T)}{3}$

(ii) $\displaystyle\int_\Omega \psi\phi_\alpha^2 \, d\mu_\Omega = \frac{\mu_\Omega(T)}{9}$



Proof : Denote the Barycentric Coordinates of a point $X \in X_\alpha + H$ w.r.t. the triangle $T_j = X_\alpha X_{\alpha_j} X_{\alpha_{j+1}}$ by (ξ,η,κ). Then the polynomial ψ can be represented in terms of the Barycentric Coordinates of X w.r.t. the triangle $T_1 = X_\alpha X_{\alpha_1} X_{\alpha_2}$ as $\psi(\xi,\eta,\kappa) = \eta^2$

It follows from the transformation matrix we have developed in Section 1.3 that the local expression of $\psi(X)$ w.r.t. the six triangles T_j are as follow :

$$\psi(\xi,\eta,\kappa) = \begin{cases} \kappa^2 & \text{in } T_2 \text{ and } T_5 \\ \eta^2 + 2\eta\kappa + \kappa^2 & \text{in } T_3 \text{ and } T_6 \\ \eta^2 & \text{in } T_1 \text{ and } T_4 \end{cases}$$

It follows that

$$\int_\Omega \psi\phi_\alpha \, d\mu_\Omega = 4\mu_\Omega(T)\int_T \xi(\eta^2 + \kappa^2 + \eta\kappa) \, d\mu_T$$

$$= 4\mu_\Omega(T)\left(\frac{2!2!}{5!} + \frac{2!2!}{5!} + \frac{2!}{5!}\right)$$

$$= \frac{\mu_\Omega(T)}{3}$$

Similarly, we have

$$\int_\Omega \psi\phi_\alpha^2 \, d\mu_\Omega = 4\mu_\Omega(T)\int_T \xi^2(\eta^2 + \kappa^2 + \eta\kappa) \, d\mu_T = \frac{\mu_\Omega(T)}{9}$$

completing the proof.

From the result of Lemma 2.6.1., we observe that the two numerical quadratures $\tilde{F}_\alpha = 2\mu_\Omega(T)$ and $\tilde{Q}_\alpha = \mu_\Omega(T)$ are not exact for all polynomials of degree 2.

Another numerical quadrature for F_α and Q_α exact for polynomials of higher degree can be derived as follows :

Assume the numerical quadrature for F_α has the form :

$$\tilde{F}_\alpha(X_\beta) = \begin{cases} af(X_\beta) & \text{if } \alpha=\beta \\ bf(X_\beta) & \text{if } |\alpha-\beta|=1 \\ 0 & \text{otherwise} \end{cases}$$

Since $\tilde{F}_\alpha$ has two parameters a and b to be determined, and the one parameter numerical quadrature is exact for all polynomials of degree 1, we may require the 7-point formula to be exact for all polynomials of degree 2.

If this is the case, we should have $F_\alpha - \tilde{F}_\alpha = 0$ for f equal to 1, and the quadrature polynomial ψ as defined in Lemma 2.6.1., that is

$$\begin{cases} \int_\Omega \phi_\alpha \, d\mu_\Omega - a - 6b = 0 \\ \\ \int_\Omega \psi\phi_\alpha \, d\mu_\Omega - 4b = 0 \end{cases} \tag{2.6.1}$$

It follows from the result of Lemma 2.6.1. that

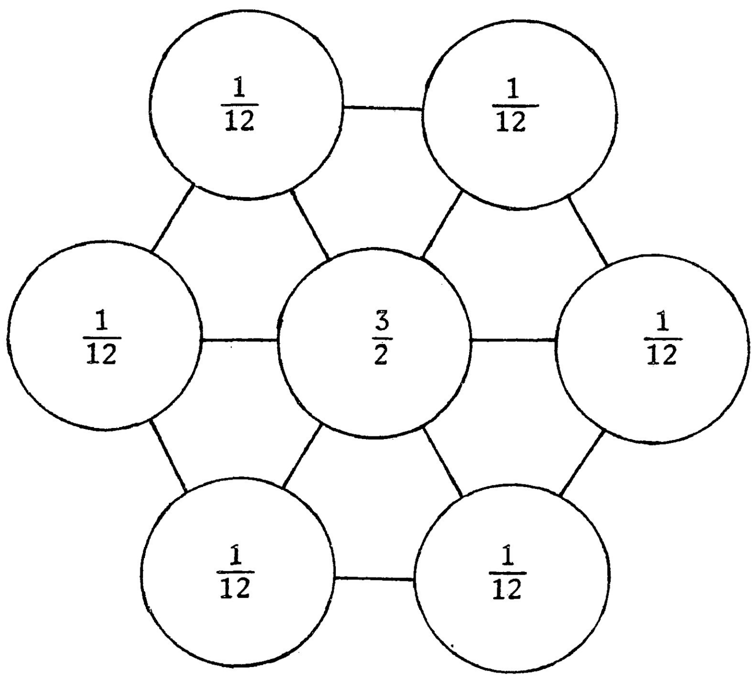
$$b = \frac{\mu_\Omega(T)}{12}$$

By substituting this into (2.6.1), we have

$$a = \frac{3}{2}\mu_\Omega(T)$$

Expressed diagrammatically, the 7-point numerical quadrature can be represented as

$\frac{\tilde{F}_\alpha}{\mu_\Omega(T)}$:



(2.6.2)

The set $B^n_{T_1} = \{ \xi^{s_1}\eta^{s_2}\kappa^{s_3} : s_i$ are non-negative integers and $\sum_i s_i = n \}$ form a basis for all homogeneous polynomials of degree n on T_1, and these polynomials can be extended to Ω in a consistent way. It is not hard to verify that the 7-point formula for F_α

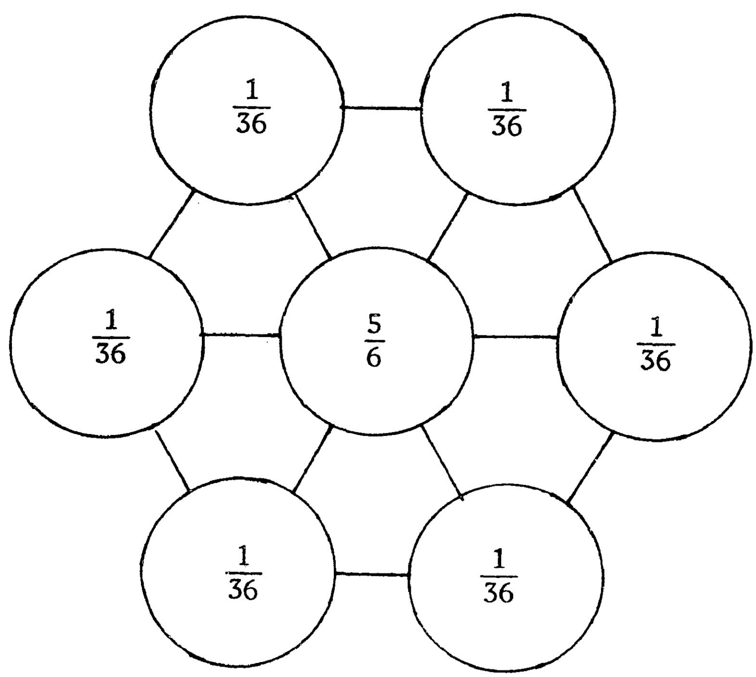
is exact for all polynomials in $B^2_{T_1}$. Since elements of $B^1_{T_1}$, $B^3_{T_1}$ are all odd functions, by the symmetric form of the integral F_α and the numerical quadrature $\tilde{F}_\alpha$, the 7-point numerical quadrature $\tilde{F}_\alpha$ is exact for all polynomials in $B^1_{T_1}$ and $B^3_{T_1}$. Thus the 7-point formula $\tilde{F}_\alpha$ is exact for all polynomials of degree ≤ 3.

Similarly, the 7-point numerical quadrature for Q_α can be obtained by solving the following system of linear equations :

$$\begin{cases} \int_\Omega \phi_\alpha^2 \, d\mu_\Omega - a - 6b = 0 \\ \int_\Omega \psi\phi_\alpha^2 \, d\mu_\Omega - 4b = 0 \end{cases}$$

and this reduces to

$\dfrac{\tilde{Q}_\alpha}{\mu_\Omega(T)}$:



It is easy to verify that the 7-point formula $\tilde{Q}_\alpha$ is consistent i.e. $Q_\alpha - \tilde{Q}_\alpha = 0$ for all $q \in P^2(\Omega)$, by applying the symmetry arguments, we conclude that $\tilde{Q}_\alpha$ is exact for all polynomials of degree ≤ 3.

If we return our attention to the integrals $\int_{T_\alpha \cup T_\beta} p \, d\mu_\Omega$ and $\int_\Omega q\phi_\alpha\phi_\beta \, d\mu_\Omega$, we observe that $Q_{\alpha,\beta}$ has support over the two adjacent triangles T_α and T_β, and for the integral $P_{\alpha,\beta}$ we only have to integrate p over the triangles T_α and T_β.

The simplest numerical quadrature for $Q_{\alpha,\beta}$ is the following 2-point formula

$$\tilde{Q}_{\alpha,\beta}(X_\gamma) = \begin{cases} aq(X_\gamma) & \text{if } \gamma=\alpha \text{ or } \beta \\ \\ 0 & \text{elsewhere} \end{cases}$$

To determine a, we may require $Q_{\alpha,\beta} - \tilde{Q}_{\alpha,\beta} = 0$ for $q = 1$, that is

$$\int_\Omega \phi_\alpha\phi_\beta \, d\mu_\Omega - 2a = 0$$

this reduces to

$$a = \frac{\mu_\Omega(T)}{12}$$

It is easy to check that the 2-point formula for $Q_{\alpha,\beta}$ is exact for all polynomials of degree ≤ 1.

Similarly, the 2-point formula for the integral $P_{\alpha,\beta}$ is

$$\tilde{P}_{\alpha,\beta}(X_\gamma) = \begin{cases} \mu_\Omega(T) & \text{if } \gamma = \alpha \text{ or } \beta \\ 0 & \text{elsewhere} \end{cases}$$

It is easy to verify that the 2-point formula $\tilde{P}_{\alpha,\beta}$ is exact for all polynomials of degree ≤ 1.

Numerical quadrature for $P_{\alpha,\beta}$ and $Q_{\alpha,\beta}$ exact for polynomials of higher degree can be obtained by putting weights at several points on the triangles T_α and T_β (see T. H. Lim [L1]).

CHAPTER 3

ERROR ANALYSIS

3.1 INTRODUCTION

Error bounds for the finite element method for elliptic boundary value problems are frequently of the form

$\| u-u^h \|_a \leq kh^s \|u\|_{k,\Omega}$, where k is a constant independent of h, the mesh parameter. In this chapter, we apply a triangular version of the Peano-Sard Kernel Theorem, proved by Frederickson [F6], to construct some kernels for the error functions $u-u_I$ and $D_i(u-u_I)$ in the Barycentric Coordinates system. Error bounds are computed from these kernels and applied to the finite element analysis of elliptic boundary value problems, to obtain an upper bound for the constant k. The expression of norms in the interpolation error bounds are simplified by an application of the generalized Hardy inequality proved by P. Frederickson and W. Eames [F5], to the norm of the form $\Big\| \|u\|_{L^1(T_i)} \Big\|_{L^2(T)}$, where T_i is a sub-triangle of T.

Barnhill and Gregory ([B1],[B2]) have applied the Sard Kernel Theorem in the rectangular coordinate system to obtain an error bound for the constant k, but their computation involves line integrals and is more complicated than the results we have obtained.

In Section 3.4, Peano-Sard Kernels for the 1-point and

7-point numerical quadratures are derived, and the error bounds for these numerical quadratures are estimated. The quadrature errors introduced by computing $\tilde{u}^h$ rather than u^h are also discussed in this section.

3.2 ERROR BOUNDS FOR INTERPOLATION ON TRIANGLES

Denote by $E(u,X) = u(X) - u_I(X)$ the error of u at $X \in \Omega$, where u_I is an interpolant of u. In particular, if u_I is a piecewise linear interpolation of u, then we have the following Theorem.

<u>Theorem 3.2.1</u> If $u \in H^2(\Omega)$, then

$$\| E(u,\cdot) \|_{L^2(\Omega)} \leq \frac{h^2}{\sqrt{17.5}} |u|_{2,\Omega} \qquad (3.2.1)$$

To prove the Theorem, we need some auxiliary lemmas and the following generalized Hardy inequality.

<u>Generalized Hardy Inequality</u> : For any $u \in L^p(T)$, $p > 1$, define

Φ by $\Phi(X) = \dfrac{\int_{T_X} u(\S) d\mu_T(\S)}{\mu_T(T_X)}$, where T is the triangle $A_0A_1A_2$ and T_X is the triangle XA_1A_2.

Then $\Phi \in L^p(T)$, and

$$\| \Phi \|_{L^p(T)} \leq \frac{2p}{p-1} \| u \|_{L^p(T)},$$

A proof of the inequality has been given by Frederickson and Eames [F5].

Lemma 3.2.1. If the error functional $E(u,X)$ is expressed in terms of the kernels in equation (1.7.14) as

$$E(u,X) = \sum_i \int_{T_i} [\tfrac{1}{2} x_i x_{i+1} D_{i,i-1} u(\S) + \tfrac{1}{2} x_i x_{i-1} D_{i,i+1} u(\S)] d\mu_{T_i}(\S) \tag{3.2.2}$$

then

$$\| E(u,\cdot) \|_{L^2(T)} \leq \frac{1}{\sqrt{120}} \sum_i (\| D_{i,i-1} u \|_{L^2(T)} + \| D_{i,i+1} u \|_{L^2(T)}) \tag{3.2.3}$$

Proof : Since $\mu_T(T_i) = x_i$, the equation (3.2.2) can be written as

$$E(u,X) = \sum_i \int_T [\kappa_{i,i-1}(\S) D_{i,i-1} u(\S) + \kappa_{i,i+1}(\S) D_{i,i+1} u(\S)] d\mu_T(\S) \tag{3.2.4}$$

where

$$\kappa_{i,i\pm 1}(\S) = \begin{cases} \frac{1}{2} x_{i\mp 1} & \text{if } \S \in T_i \\ 0 & \text{otherwise} \end{cases}$$

By applying the triangle inequality and the Cauchy-Schwarz inequality to the equation (3.2.4), we have

$$|E(u,X)| \le \sum_i (\| \kappa_{i,i-1} \|_{L^2(T)} \|D_{i,i-1} u\|_{L^2(T)} + \| \kappa_{i,i+1} \|_{L^2(T)} \|D_{i,i+1} u \|_{L^2(T)} \tag{3.2.5}$$

Since the kernels $\kappa_{i,i\pm 1}(X)$ vanish outside the triangle T_i, we have

$$\begin{aligned} \|\kappa_{i,i\pm 1}\|^2_{L^2(T)} &= \int_T \kappa_{i,i\pm 1}(X)\, d\mu_T(X) \\ &= \frac{1}{4} x^2_{i\mp 1} \mu_T(T_i) \\ &= \frac{1}{4} x^2_{i\mp 1} x_i \end{aligned}$$

Substituting this into (3.2.5), followed by taking the L^2 norm of $E(u,X)$ over T, together with the application of the triangle inequality and the Cauchy-Schwarz inequality, we get

$$\begin{aligned} \|E(u,\cdot)\|_{L^2(T)} \le \sum_i [(\int \frac{1}{4} x^2_{i+1} x_i\, d\mu_T(X))^{\frac{1}{2}} \|D_{i,i-1} u\|_{L^2(T)} + \\ (\int \frac{1}{4} x^2_{i-1} x_i\, d\mu_T(X))^{\frac{1}{2}} \|D_{i,i+1} u\|_{L^2(T)} \end{aligned}$$

$$= \frac{1}{\sqrt{120}} \sum_i (\|D_{i,i-1}u\|_{L^2(T)} + \|D_{i,i+1}u\|_{L^2(T)})$$

completing the proof.

Lemma 3.2.2. If the error function $E(u,X)$ is expressed in terms of the kernels in equation (1.7.16) as

$$E(u,X) = \sum_i \int_{T_i} (-\tfrac{1}{2}x_{i-1}x_{i+1}) D_{i,i}u(\S)\, d\mu_{T_i}(\S) \qquad (3.2.6)$$

then

$$\|E(u,\cdot)\|_{L^2(T)} \leq \frac{2}{\sqrt{90}} \sum_i \|D_{i,i}u\|_{L^2(T)} \qquad (3.2.7)$$

Proof : It follows from (3.2.6) that

$$|E(u,X)| \leq \sum_i \tfrac{1}{2}x_{i-1}x_{i+1} \|D_{i,i}u\|_{L^1(T_i)}$$

By taking the L^2 norm of $E(u,X)$ over the triangle T, together with the application of the triangle inequality and the Cauchy-Schwarz inequality, we have

$$\|E(u,\cdot)\|_{L^2(T)} \leq \sum_i \left(\int_T \tfrac{1}{4}x_{i-1}^2 x_{i+1}^2\, d\mu_T(X)\right)^{\frac{1}{2}} \left\| \|D_{i,i}u\|_{L^1(T_i)} \right\|_{L^2(T)}$$

$$= \frac{1}{\sqrt{360}} \sum_i \left\| \|D_{i,i}u\|_{L^1(T_i)} \right\|_{L^2(T)}$$

Applying the generalized Hardy inequality to the norm $\left\| \| D_{i,i} u \|_{L^1(T_i)} \right\|_{L^2(T)}$, we have

$$\| E(u,\cdot) \|_{L^2(T)} \le \frac{2}{\sqrt{90}} \sum_i \| D_{i,i} u \|_{L^2(T)}$$

completing the proof.

Remark: The reader may wonder why we use two different techniques to prove the Lemma 3.2.1 and Lemma 3.2.2. This is because the L^2 norm of the kernal $\kappa_{i,i}$ in (3.2.6) is

$$\| \kappa_{i,i} \|_{L^2(T)} = \frac{1}{2} x_{i-1} x_{i+1} x_i^{-\frac{1}{2}}$$

and the L^2 norm of $\frac{1}{2} x_{i-1} x_{i+1} x_i^{-\frac{1}{2}}$ does not exists. Thus we cannot apply the Cauchy-Schwarz inequality to obtain a L^2 error bound for $E(u,X)$.

However, the technique for proving the Lemma 3.2.2 can be applied to Lemma 3.2.1., but the result will be

$$\| E(u,\cdot) \|_{L^2(T)} \le \frac{2}{\sqrt{90}} \sum_i \left(\| D_{i,i-1} u \|_{L^2(T)} + \| D_{i,i+1} u \|_{L^2(T)} \right)$$

that is, a larger error bound is obtained.

Since the kernels for the error functional $E(u,X)$ are not unique, as we have proved in Lemma 3.2.1 and Lemma 3.2.2, different kernels may end up with a different upper error bound. Now we shall combine the results of Lemma 3.2.1 and Lemma 3.2.2 to prove the Theorem 3.2.1.

Proof of Theorem 3.2.1.

It follows from the inequality (3.2.3) that

$$120\|E(u,\cdot)\|^2_{L^2(T)} \leq [\sum_i(\|D_{i,i-1}u\|_{L^2(T)}+\|D_{i,i+1}u\|_{L^2(T)})]^2$$

$$\leq 12(\|D_{01}u\|^2_{L^2(T)}+\|D_{12}u\|^2_{L^2(T)}+\|D_{20}u\|^2_{L^2(T)})$$

From (3.2.7) we have

$$\frac{45}{2}\| E(u,\cdot)\|^2_{L^2(T)} \leq (\sum_i\|D_{i,i}u\|_{L^2(T)})^2 \leq 3\sum_i\|D_{i,i}u\|^2_{L^2(T)}$$

The above two inequalities follow from the fact that

$$(\sum_{i=1}^n a_i)^2 \leq n\sum_{i=1}^n a_i^2 \qquad \text{for all } a_i \in R .$$

It follows that

$$(10+\frac{15}{2})\|E(u,\cdot)\|^2_{L^2(T)} \leq \sum_{|\alpha|=2}\|D^\alpha u\|^2_{L^2(T)} = h^4|u|^2_{2,T}.$$

this reduces to

$$\|E(u,\cdot)\|_{L^2(\Omega)} = \left(\sum_{T\in\tau_h} \mu_\Omega(T)\|E(u,\cdot)\|^2_{L^2(T)}\right)^{\frac{1}{2}}$$

$$\leq \frac{h^2}{\sqrt{17.5}}\left(\sum_{T\in\tau_h} \mu_\Omega(T)\,|u|^2_{2,T}\right)^{\frac{1}{2}}$$

$$= \frac{h^2}{\sqrt{17.5}}|u|_{2,\Omega}$$

completing the proof.

To obtain an error bound for the energy norm $\| u-u_I \|_\Delta$, we have the following Theorem.

<u>Theorem 3.2.2.</u> If $u \in H^3(\Omega)$ and u_I is a piecewise linear interpolation to u, then

$$\| u-u_I \|_\Delta \leq \left(\frac{763}{1080}\right)^{\frac{1}{2}} h \| u \|_{3,\Omega}$$

for sufficiently small h $(|h|\leq 1)$.

To prove the Theorem, we need the following two lemmas.

<u>Lemma 3.2.1.</u> If $u \in H^2(\Omega)$ and u_I is a piecewise linear

interpolation of u, then the error of the derivative $D_iE(u,X) = D_i(u(X)-u_I(X))$, $X \in T = A_iA_{i+1}A_{i-1}$ has a representation of the form

$$\text{(a)}\quad D_iE(u,X) = \frac{1}{2}\int_{A_{i-1}}^{X} \S_i^{T_i}[x_iD_{i+1,i}u(\S)-x_{i+1}D_{i,i}u(\S)]\,d\S -$$

$$\frac{1}{2}\int_{X}^{A_{i+1}} \S_i^{T_i}[x_iD_{i-1,i}u(\S)-x_{i-1}D_{i,i}u(\S)]\,d\S +$$

$$\frac{1}{4}\int_{T_i}[x_i(D_{i+1,i}u(\S)-D_{i-1,i}u(\S))-(x_{i+1}-x_{i-1})D_{i,i}u(\S)]d\mu_{T_i}(\S)$$

where $\S_i^{T_i}$ is the first Barycentric Coordinate of $\S$ w.r.t. $T_i = XA_{i+1}A_{i-1}$

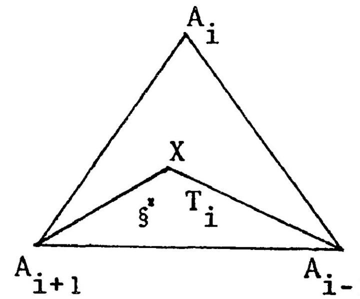
(b) In addition to that, if $u \in H^3(\Omega)$, then $D_iE(u,X)$ can be represented in terms of surface integrals of derivatives up to order 3 as

$$D_iE(u,X) = \frac{1}{2}\int_{T_i}[x_i(D_{i+1,i}u(\S)-D_{i-1,i}u(\S))-(x_{i+1}-x_{i-1})D_{i,i}u(\S) -$$

$$\S_i^{T_i}(x_i^2D_{012}u(\S)-x_ix_{i+1}D_{i-1,i,i}u(\S)-x_{i-1}x_iD_{i+1,i,i}u(\S) +$$

$$x_{i-1}x_{i+1}D_{i,i,i}u(\S))]d\mu_{T_i}(\S)$$

<u>Proof</u> : Denote by $D_{i+1}^{T_i}\cdot$ and $D_{i-1}^{T_i}\cdot$ the two normalized derivatives $D_{A_{i-1},X}\cdot$ and $D_{XA_{i+1}}\cdot$ respectively.



Then

$$D_iE(u,X) = D_i[u(X) - \sum_j x_j u(A_j)]$$

$$= D_i u(X) - [u(A_{i-1}) - u(A_{i+1})]$$

$$= \frac{1}{2}[\int_{A_{i-1}}^{X} D_{i+1}^{T_i}(\S_i^{T_i} D_i u(\S))\, d\S -$$

$$\int_X^{A_{i+1}} D_{i-1}^{T_i}(\S_i^{T_i} D_i u(\S))\, d\S] - \int_{A_{i+1}}^{A_{i-1}} D_i u(\S)\, d\S$$

$$= \frac{1}{2}\int_{A_{i-1}}^{X} \S_i^{T_i} D_{i+1}^{T_i}(D_i u(\S))d\S - \frac{1}{2}\int_X^{A_{i+1}} \S_i^{T_i} D_{i-1}^{T_i}(D_i u(\S))\, d\S -$$

$$\frac{1}{2}(\int_{A_{i+1}}^{A_{i-1}} D_i u(\S)\, d\S - \int_{A_{i-1}}^{X} D_i u(\S)\, d\S) + \frac{1}{2}(\int_X^{A_{i+1}} D_i u(\S)\, d\S -$$

$$\int_{A_{i+1}}^{A_{i-1}} D_i u(\S)\, d\S)$$

$$= \frac{1}{2}\int_{A_{i-1}}^{X} \S_i^{T_i} D_{i+1}^{T_i}(D_i u(\S))\, d\S - \frac{1}{2}\int_{X}^{A_{i+1}} \S_i^{T_i} D_{i-1}^{T_i}(D_i u(\S))\, d\S -$$

$$\frac{1}{4}\int_{T_i} D_{i-1}^{T_i}(D_i u(\S))\, d\mu_{T_i}(\S) + \frac{1}{4}\int_{T_i} D_{i+1}^{T_i}(D_i u(\S))\, d\mu_{T_i}(\S) \tag{3.2.8}$$

From (1.7.7) we have

$$D_{i\pm 1}^{T_i}(D_i u(\S)) = x_i D_{i\pm 1,i} u(\S) - x_{i\pm 1} D_{i,i} u(\S) \tag{3.2.9}$$

substituting this into (3.2.8), we obtain

$$D_i E(u,X) = \frac{1}{2}\int_{A_{i-1}}^{X} \S_i^{T_i}(x_i D_{i+1,i} u(\S) - x_{i+1} D_{i,i} u(\S))\, d\S -$$

$$\frac{1}{2}\int_{X}^{A_{i+1}} \S_i^{T_i}(x_i D_{i-1,i} u(\S) - x_{i-1} D_{i,i} u(\S))\, d\S -$$

$$\frac{1}{4}\int_{T_i} (x_i D_{i-1,i} u(\S) - x_{i-1} D_{i,i} u(\S))\, d\mu_{T_i}(\S) +$$

$$\frac{1}{4}\int_{T_i} (x_i D_{i+1,i} u(\S) - x_{i+1} D_{i,i} u(\S))\, d\mu_{T_i}(\S)$$

$$= \frac{1}{2}\int_{A_{i-1}}^{X} §_i^{T_i}(x_i D_{i+1,i}u(§) - x_{i+1}D_{i,i}u(§))\, d§ -$$

$$\frac{1}{2}\int_{X}^{A_{i+1}} §_i^{T_i}(x_i D_{i-1,i}u(§) - x_{i-1}D_{i,i}u(§))\, d§ +$$

$$\frac{1}{4}\int_{T_i} [x_i(D_{i+1,i}u(§) - D_{i-1,i}u(§)) - (x_{i+1} - x_{i-1})D_{i,i}u(§)]d\mu_{T_i}(§)$$

We have proved the part (a).

Since $§_i^{T_i}$ vanishes on the side $A_{i+1}A_{i-1}$, it follows from (3.2.8) that

$$D_i E(u,X) = \frac{1}{2}\Big[\int_{A_{i-1}}^{X} §_i^{T_i} D_{i+1}^{T_i}(D_i u(§))\, d§ - \int_{A_{i+1}}^{A_{i-1}} §_i^{T_i} D_{i+1}^{T_i}(D_i u(§)) d§\Big] -$$

$$\frac{1}{2}\Big[\int_{X}^{A_{i+1}} §_i^{T_i} D_{i-1}^{T_i}(D_i u(§))\, d§ - \int_{A_{i+1}}^{A_{i-1}} §_i^{T_i} D_{i-1}^{T_i}(D_i u(§))\, d§\Big] +$$

$$\frac{1}{4}\int_{T_i} [D_{i+1}^{T_i}(D_i u(§)) - D_{i-1}^{T_i}(D_i u(§))]\, d\mu_{T_i}(§).$$

If $u \in H^3(\Omega)$, then by Lemma 1.4.3, we get

$$D_iE(u,X) = -\frac{1}{4}\int_{T_i} D_{i-1}^{T_i}[\S_i^{T_i} D_{i+1}^{T_i}(D_i u(\S))]\ d\mu_{T_i}(\S) -$$

$$\frac{1}{4}\int_{T_i} D_{i+1}^{T_i}[\S_i^{T_i} D_{i-1}^{T_i}(D_i u(\S))]\ d\mu_{T_i}(\S) +$$

$$\frac{1}{4}\int_{T_i} [D_{i+1}^{T_i}(D_i u(\S)) - D_{i-1}^{T_i}(D_i u(\S))] d\mu_{T_i}(\S)$$

$$= \frac{1}{2}\int_{T_i} [D_{i+1}^{T_i}(D_i u(\S)) - D_{i-1}^{T_i}(D_i u(\S)) - \S_i^{T_i} D_{i-1}^{T_i}(D_{i+1}^{T_i}(D_i u(\S)))] d\mu_{T_i}(\S) \qquad (3.2.10)$$

From (3.2.9), we have

$$D_{i-1}^{T_i}[D_{i+1}^{T_i}(D_i u(\S))] = D_{i-1}^{T_i}(x_i D_{i+1,i} u(\S) - x_{i+1} D_{i,i} u(\S))$$

$$= x_i^2 D_{012} u(\S) - x_i x_{i+1} D_{i-1,i,i} u(\S) - x_{i-1} x_i D_{i+1,i,i} u(\S) +$$

$$x_{i-1} x_{i+1} D_{i,i,i} u(\S)$$

substituting this into (3.2.10), we get

$$D_iE(u,X) = \frac{1}{2}\int_{T_i} [x_i(D_{i+1,i} u(\S) - D_{i-1,i} u(\S)) - (x_{i+1} - x_{i-1}) D_{i,i} u(\S) -$$

$$\S_i^{T_i}(x_i^2 D_{012} u(\S) - x_i x_{i+1} D_{i-1,i,i} u(\S) - x_{i-1} x_i D_{i+1,i,i} u(\S) +$$

$$x_{i-1} x_{i+1} D_{i,i,i} u(\S))]\ d\mu_{T_i}(\S)$$

completing the proof.

<u>Lemma 3.2.2</u> If $u \in H^3(\Omega)$ and u_I is a piecewise linear interpolation of u, then

$$\|D_i(u-u_I)\|_{L^2(T)} \le \frac{1}{\sqrt{12}}(\|D_{i+1,i}u\|_{L^2(T)}+\|D_{i-1,i}u\|_{L^2(T)}) + \frac{1}{\sqrt{240}}\|D_{012}u\|_{L^2(T)} + \frac{1}{\sqrt{720}}(\|D_{i-1,i,i}u\|_{L^2(T)}+\|D_{i+1,i,i}u\|_{L^2(T)}) + \frac{2}{\sqrt{6}}\|D_{i,i}u\|_{L^2(T)}+\frac{2}{\sqrt{90}}\|D_{i,i,i}u\|_{L^2(T)} \tag{3.2.11}$$

<u>Proof</u>: From Lemma 3.2.1. we have

$$D_i(u(X)-u_I(X)) = \frac{1}{2}\int_T \chi_{T_i}(\S)(D_{i+1,i}u(\S)-D_{i-1,i}u(\S)-\S_i^{T_i}(x_iD_{012}u(\S) - x_{i+1}D_{i-1,i,i}u(\S)-x_{i-1}D_{i+1,i,i}u(\S)))\, d\mu_T(\S) - \frac{1}{2}\int_{T_i}[(x_{i+1}-x_{i-1})D_{i,i}u(\S)+\S_i^{T_i}x_{i-1}x_{i+1}D_{i,i,i}u(\S)]\, d\mu_{T_i}(\S)$$

where χ_{T_i} denotes the characteristic function of T_i.

By applying the triangle inequality and the Cauchy-Schwarz inequality to the above equation, we have

$$|D_i(u-u_I)| \le \frac{1}{2}x_i^{\frac{1}{2}}(\| D_{i+1,i}u\|_{L^2(T)} + \| D_{i-1,i}u\|_{L^2(T)} +$$

$$\frac{x_i}{\sqrt{6}}\|D_{012}u\|_{L^2(T)} + \frac{x_{i+1}}{\sqrt{6}}\| D_{i-1,i,i}u\|_{L^2(T)} +$$

$$\frac{x_{i-1}}{\sqrt{6}}\| D_{i+1,i,i}u\|_{L^2(T)}) + \frac{1}{2}|x_{i+1}-x_{i-1}| \ \|D_{i,i}u\|_{L^1(T_i)} +$$

$$\frac{1}{2}x_{i-1}x_{i+1}\|D_{i,i,i}u\|_{L^1(T_i)}$$

by taking the L^2 norm of $D_i(u-u_I)$ over the triangle T, together with the application of the triangle inequality and the Cauchy-Schwarz inequality, we obtain

$$\| D_i(u-u_I)\|_{L^2(T)} \le \frac{1}{\sqrt{12}}(\| D_{i+1,i}u\|_{L^2(T)} + \| D_{i-1,i}u\|_{L^2(T)}) +$$

$$\frac{1}{\sqrt{240}}\| D_{012}u\|_{L^2(T)} + \frac{1}{\sqrt{720}}(\| D_{i-1,i,i}u\|_{L^2(T)} +$$

$$\| D_{i+1,i,i}u\|_{L^2(T)}) + \frac{1}{\sqrt{24}}\Big\| \ \|D_{i,i}u\|_{L^1(T_i)} \Big\|_{L^2(T)} +$$

$$\frac{1}{\sqrt{360}}\Big\| \ \| D_{i,i,i}u\|_{L^1(T_i)} \Big\|_{L^2(T)}$$

applying the Generalized Hardy inequality to the norm

$\Big\| \ \| \cdot \|_{L^1(T_i)} \Big\|_{L^2(T)}$, we have

$$\| D_i(u-u_I) \|_{L^2(T)} \le \frac{1}{\sqrt{12}}(\| D_{i+1,i}u\|_{L^2(T)} + \| D_{i-1,i}u\|_{L^2(T)}) + \frac{1}{\sqrt{240}}\| D_{012}u\|_{L^2(T)} + \frac{1}{\sqrt{720}}(\| D_{i-1,i,i}u\|_{L^2(T)} + \| D_{i+1,i,i}u\|_{L^2(T)}) + \frac{2}{\sqrt{6}}\| D_{i,i}u\|_{L^2(T)} + \frac{2}{\sqrt{90}}\| D_{i,i,i}u\|_{L^2(T)}.$$

completing the proof.

<u>Proof of Theorem 3.2.2.</u>

$$\|u-u_I\|_\Delta^2 = \sum_{T\in\tau^h} \mu_\Omega(T)\int_T \frac{2}{3h^2}\sum_i [D_i(u-u_I)]^2 \, d\mu_T$$

$$= \sum_{T\in\tau^h} \frac{2\mu_\Omega(T)}{3h^2}\sum_i \| D_i(u-u_I)\|^2_{L^2(T)} \qquad (3.2.12)$$

By applying the Cauchy-Schwarz inequality to the right hand side of the inequality (3.2.11), we get

$$\| D_i(u-u_I) \|_{L^2(T)} \le (\tfrac{1}{6}+\tfrac{1}{6}+\tfrac{1}{80}+\tfrac{1}{720}+\tfrac{1}{720}+\tfrac{4}{6}+\tfrac{4}{90})^{\frac{1}{2}}(\tfrac{1}{2}\|D_{i+1}u\|^2_{L^2(T)} + \tfrac{1}{2}\|D_{i-1}u\|^2_{L^2(T)} + \|D_{i,i}u\|^2_{L^2(T)} + \tfrac{1}{3}\|D_{012}u\|^2_{L^2(T)} + \|D_{i+1,i,i}u\|^2_{L^2(T)} + \|D_{i-1,i,i}u\|^2_{L^2(T)} + \|D_{i,i,i}u\|^2_{L^2(T)})^{\frac{1}{2}}$$

It follows that

$$\sum_i \|D_i(u-u_I)\|^2_{L^2(T)} \le \frac{763}{720}\{\|D_{01}u\|^2_{L^2(T)} + \|D_{12}u\|^2_{L^2(T)} + \|D_{20}u\|^2_{L^2(T)} + \|D_{00}u\|^2_{L^2(T)} + \|D_{11}u\|^2_{L^2(T)} + \|D_{22}u\|^2_{L^2(T)} + \|D_{012}u\|^2_{L^2(T)} + \sum_i(\|D_{i+1,i,i}u\|^2_{L^2(T)} + \|D_{i-1,i,i}u\|^2_{L^2(T)} + \|D_{i,i,i}u\|^2_{L^2(T)})\}$$

$$\le \frac{763}{720}h^4(|u|^2_{2,T} + h^2|u|^2_{3,T})$$

$$\le \frac{763}{720}h^4\|u\|^2_{3,T} \qquad \text{for sufficiently small } h \ (|h|\le 1).$$

Substituting this into (3.2.12), we have

$$\|u-u_I\|_\Delta \le (\frac{763}{1080})^{\frac{1}{2}} h(\sum_{T\in\tau^h} \mu_\Omega(T)\|u\|^2_{3,T})^{\frac{1}{2}} = (\frac{763}{1080})^{\frac{1}{2}} h\|u\|_{3,\Omega}$$

completing the proof.

3.3 ERROR BOUNDS OF THE RITZ APPROXIMATION

As we have discussed in Section 2.3, the energy norm $\|u\|_a = (\int_\Omega (p\nabla u \cdot \nabla u + qu^2)\, d\mu_\Omega)^{\frac{1}{2}}$ is equivalent to the Sobolev norm $\|u\|_{1,\Omega}$, and provides a means of measuring how close the Ritz approximation u^h is to the true solution u.

The following Theorem [S4,p.39] is fundamental to the Ritz theorey.

<u>Theorem 3.3.1.</u> [S3] If the function u minimizes $I(v)$ over the admissible space H_g and $S_g = S_0 + g$ is a closed affine subspace of H_g, then

(a) $a(u-u^h, u-u^h) = \min_{v^h \in S_g} a(u-v^h, u-v^h)$ (3.3.1)

(b) $a(u-u^h, v^h) = 0$ for all $v^h \in S_0$ (3.3.2)

(c) $a(u^h, v^h) = (f, v^h)$ for all $v^h \in S_0$ (3.3.3)

In particular, if $S_g = H_g$, then

$a(u,v) = (f,v)$ for all $v \in H_0$ (3.3.4)

<u>Corollary 3.3.1.</u> [S3] It follows from (3.3.2) that $a(u-u^h, u^h-g) = 0$ and $a(u-u^h, u-u^h) = a(u-g, u-g) - a(u^h-g, u^h-g)$. Furthermore, since $a(u-u^h, u-u^h) \geq 0$, the strain energy in $u^h - g$ always underestimates the strain energy in $u - g$, that is $a(u^h-g, u^h-g) \leq a(u-g, u-g)$.

Corollary 3.3.2. Let u_I be an interpolant of u in S_g, then

$$a(u-u^h,u-u^h) \le a(u-u_I,u-u_I) \tag{3.3.5}$$

In fact $a(u-u^h,u-u^h) + a(u^h-u_I,u^h-u_I) = a(u-u_I,u-u_I)$ (3.3.6)

Proof: Inequality (3.3.5) follows directly from equation (3.3.1).

$$a(u-u_I,u-u_I) = a(u-u^h+u^h-u_I,u-u^h+u^h-u_I)$$

$$= a(u-u^h,u-u^h)+2a(u-u^h,u^h-u_I)+a(u^h-u_I,u^h-u_I)$$

since $u^h-u_I \in S_0$, from (3.3.2), we have

$$a(u-u^h,u^h-u_I) = 0$$

which implies

$$a(u-u^h,u-u^h)+a(u^h-u_I,u^h-u_I) = a(u-u_I,u-u_I)$$

completing the proof.

To obtain an error bound for the energy norm $\|u-u^h\|_a$, we have the following Theorem :

Theorem 3.3.2. If $u \in H^3(\Omega)$ and u_I is a piecewise linear interpolant to u in $S^{1,0}$, then

$$\|u-u^h\|_a \le h \max\{(\tfrac{763}{1080})^{\frac{1}{2}}\|p\|_\infty^{\frac{1}{2}}, \frac{h\|q\|_\infty^{\frac{1}{2}}}{\sqrt{17.5}}\}\|u\|_{3,\Omega}$$

Proof:

$$\| u-u_I \|_a = \{\int_\Omega [p\nabla(u-u_I)\cdot\nabla(u-u_I)+q(u-u_I)^2]\ d\mu_\Omega\ \}^{\frac{1}{2}}$$

$$\leq [\| p \|_\infty \int_\Omega \nabla(u-u_I)\cdot\nabla(u-u_I)\ d\mu_\Omega + \| q \|_\infty \| u-u_I \|^2_{L^2(\Omega)}]^{\frac{1}{2}}$$

$$\leq \| p \|_\infty^{\frac{1}{2}} \| u-u_I \|_\Delta + \| q \|_\infty^{\frac{1}{2}} \| u-u_I \|_{L^2(\Omega)} \qquad (3.3.8)$$

The inequality (3.3.8) followed from the fact that

$$(a^2+b^2)^{\frac{1}{2}} \leq a + b \qquad \text{if } a, b \geq 0$$

From Theorem 3.2.1. and Theorem 3.2.2. we have

$$\| u-u_I \|_a \leq (\frac{763}{1080})^{\frac{1}{2}} \| p \|_\infty^{\frac{1}{2}} h \| u \|_{3,\Omega} + \frac{\| q \|_\infty^{\frac{1}{2}} h^2}{\sqrt{17.5}} |u|_{2,\Omega}$$

$$\leq h \max \{ (\frac{763}{1080})^{\frac{1}{2}} \| p \|_\infty^{\frac{1}{2}}, \frac{h \| q \|_\infty^{\frac{1}{2}}}{\sqrt{17.5}} \} \| u \|_{3,\Omega}$$

The result of the Theorem follows from Corollary 3.3.2.

It follows from Theorem 3.3.2. that the Ritz-Galerkin solution to the problem $Lu = f$ with linear element has a rate of convergence of order h in the energy norm.

3.4 QUADRATURE ERRORS AND THEIR EFFECT ON THE NUMERICAL SOLUTION OF BOUNDARY VALUE PROBLEMS

In this section, we shall derive the Peano-Sard kernel of the 1-point and 7-point numerical quadratures of the integral $\int_\Omega f\phi_\alpha d\mu_\Omega$ and obtain an error bounds for these two numerical quadratures. The effect of the quadrature errors to the solution of the boundary value problems is also discussed in this section.

For simplicity, we denote by X_0 the centre X_α of the hexagon $X_\alpha+H$ and by $X_j, j=1, \ldots 6$ the six vertices of $X_\alpha+H$.

To get an estimate for the 1-point numerical quadrature error, we have the following Theorem.

<u>Theorem 3.4.1.</u> If $f \in H^2(\Omega)$ and $\tilde{F}_\alpha(f) = 2\mu_\Omega(T)f(X_\alpha)$, then

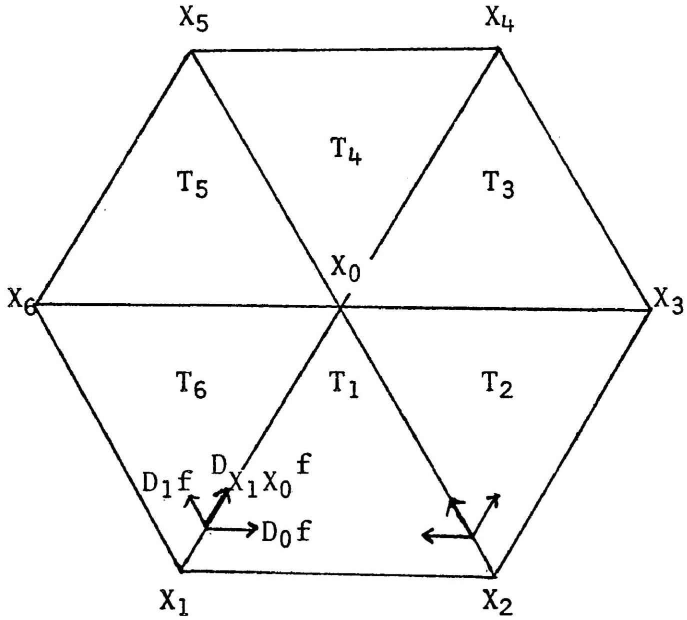
$$\Big(\sum_{\alpha\in\mathring{\Gamma}_h} |E(f,X_\alpha)|^2\Big)^{\frac{1}{2}} \le \Big(\frac{23}{560}\mu_\Omega(T)\Big)^{\frac{1}{2}} h^2 |f|_{2,\Omega} \qquad (3.4.1)$$

To prove the Theorem, we need the following two auxillary lemmas.

<u>Lemma 3.4.1.</u> If $f \in H^2(\Omega)$ and $P(\xi_0)$ is a real valued function of ξ_0 defined on each of the triangular elements $T_j = X_0X_jX_{j+1}$, then

$$\sum_{j=1}^{6}\int_{X_j}^{X_0} P(\xi_0)D_{X_jX_0}f\,dX = -\sum_{j=1}^{6}\int_{T_j}\tfrac{1}{2}P(\xi_0)D_{00}f\,d\mu_{T_j}$$

Proof:



We observe that along the side X_jX_0, $D_{X_jX_0}f$ can be decomposed into the sum of the two derivatives D_0f and D_1f. Thus, we have

$$\sum_{j=1}^{6}\int_{X_j}^{X_0} P(\xi_0)D_{X_jX_0}f\,dX = \sum_{j=1}^{6}\left[\int_{X_0}^{X_j} P(\xi_0)D_0f\,dX + \int_{X_0}^{X_j} P(\xi_0)D_1f\,dX\right]$$

since the derivative D_1f w.r.t. T_j along the side X_jX_0 is the same as the derivative $-D_0f$ w.r.t. T_{j-1}, we have

$$\sum_{j=1}^{6}\int_{X_j}^{X_0} P(\xi_0)D_{X_jX_0} f\ dX = -\sum_{j=1}^{6}[\int_{X_{j+1}}^{X_0} P(\xi_0)D_0 f\ dX - \int_{X_0}^{X_j} P(\xi_0)D_0 f\ dX]$$

$$= -\sum_{j=1}^{6}\int_{T_j} \frac{1}{2}P(\xi_0)D_{00} f\ d\mu_{T_j}$$

completing the proof.

Lemma 3.4.2. If $f \in H^2(\Omega)$, then the error of the 1-point numerical quadrature has a representation of the form :

$$E(f,X_\alpha) = \int_\Omega f\phi_\alpha\ d\mu_\Omega - 2h^2 f(X_\alpha)$$

$$= \sum_{j=1}^{6}\mu_\Omega(T)\int_{T_j}(\frac{\xi_0^2}{2} - \frac{\xi_0^3}{3} - \frac{\xi_0\xi_1\xi_2}{2})D_{00}f\ d\mu_{T_j}$$

(3.4.2)

Proof :

$$E(f,X_\alpha) = \int_\Omega f\phi_\alpha\ d\mu_\Omega - 2h^2 f(X_\alpha)$$

$$= \sum_{j=1}^{6}\mu_\Omega(T)[\int_{T_j} f\xi_0\ d\mu_{T_j} - \int_{T_j} f(X_0)\xi_0\ d\mu_{T_j}]$$

$$= \sum_{j=1}^{6}\mu_\Omega(T)\int_{T_j}[f - f(X_0)]\xi_0\ d\mu_{T_j}$$

$$= \sum_{j=1}^{6}\mu_\Omega(T)\int_{T_j}[f - f(X_0)][D_0\frac{\xi_0}{2}(\xi_2 - \xi_1)]\ d\mu_{T_j}$$

It follows from Lemma 1.4.4 and the fact ξ_1 and ξ_2 vanish on $X_{j+1}X_0$ and X_0X_j respectively that

$$E(f,X_\alpha) = \sum_{j=1}^{6} \mu_\Omega(T)[\int_{X_{j+1}}^{X_0} \xi_0(1 - \xi_0)(f - f(X_0))\, dX +$$

$$\int_{X_0}^{X_j} \xi_0(1 - \xi_0)(f - f(X_0))\, dX + \int_{T_j} D_0(\tfrac{1}{2}\xi_0\xi_1\xi_2)D_0 f\, d\mu_{T_j}]$$

$$= \sum_{j=1}^{6} \mu_\Omega(T)[2\int_{X_{j+1}}^{X_0} \xi_0(1 - \xi_0)(f-f(X_0))\, dX + \int_{X_{j+1}}^{X_0} \xi_0\xi_1\xi_2 D_0 f\, dX -$$

$$\int_{X_0}^{X_j} \xi_0\xi_1\xi_2 D_0 f\, dX - \int_{T_j} \tfrac{1}{2}\xi_0\xi_1\xi_2 D_{00} f\, d\mu_{T_j}]$$

$$= \sum_{j=1}^{6} \mu_\Omega(T)[2(\frac{\xi_0^2}{2} - \frac{\xi_0^3}{3})(f - f(X_0))\Big|_{X_j}^{X_0} - 2\int_{X_j}^{X_0} (\frac{\xi_0^2}{2} - \frac{\xi_0^3}{3})D_{X_jX_0} f\, dX -$$

$$\int_{T_j} \tfrac{1}{2}\xi_0\xi_1\xi_2 D_{00} f\, d\mu_{T_j}]$$

$$= -\mu_\Omega(T) \sum_{j=1}^{6} [\int_{X_j}^{X_0} (\xi_0^2 - \tfrac{2}{3}\xi_0^3)D_{X_jX_0} f\, dX + \int_{T_j} \tfrac{1}{2}\xi_0\xi_1\xi_2 D_{00} f\, d\mu_{T_j}]$$

It follows from Lemma 3.4.1. that

$$E(f,X_\alpha) = \sum_{j=1}^{6} \mu_\Omega(T)\int_{T_j} (\frac{\xi_0^2}{2} - \frac{\xi_0^3}{3} - \tfrac{1}{2}\xi_0\xi_1\xi_2)D_{00} f\, d\mu_{T_j}$$

completing the proof.

Proof of Theorem 3.4.1.

Application of the triangle inequality and the Cauchy-Schwarz inequality to equation (3.4.2), we get

$$|E(f,X_\alpha)| \le \sum_{j=1}^{6} \mu_\Omega(T)[\int_{T_j} (\tfrac{1}{2}\xi_0^2 - \tfrac{1}{3}\xi_0^3 - \tfrac{1}{2}\xi_0\xi_1\xi_2)^2 \, d\mu_{T_j}]^{\frac{1}{2}} \|D_{00}f\|_{L^2(T_j)}$$

$$= (\tfrac{23}{3360})^{\frac{1}{2}}\mu_\Omega(T) \sum_{j=1}^{6} \|D_{00}f\|_{L^2(T_j)}$$

It follows that

$$\sum_{\alpha\in\mathring{\Gamma}_\alpha} |E(f,X_\alpha)|^2 \le \frac{23}{3360}\mu_\Omega(T) \sum_{\alpha\in\mathring{\Gamma}_h} \mu_\Omega(T)[\sum_{j=1}^{6} \|D_{00}f\|_{L^2(T_j)})]^2$$

$$\le \frac{23}{560}\mu_\Omega(T) \sum_{\alpha\in\mathring{\Gamma}_h} [\mu_\Omega(T)\sum_{j=1}^{6} \|D_{00}f\|^2_{L^2(T_j)}]$$

We observe that for each $T = X_0X_1X_2 \in \tau^h$ and each i, the term $\|D_{i,i}f\|^2_{L^2(T_j)}$ appears in the right hand side of the above inequality at most once, thus we have

$$(\sum_{\alpha\in\mathring{\Gamma}_\alpha} |E(f,X_\alpha)|^2)^{\frac{1}{2}} \le [\frac{23}{560}\mu_\Omega(T)]^{\frac{1}{2}}(\sum_{T\in\tau^h} \mu_\Omega(T)\sum_{i=1}^{3} \|D_{i,i}f\|^2_{L^2(T)})^{\frac{1}{2}}$$

$$\le [\frac{23}{560}\mu_\Omega(T)]^{\frac{1}{2}} h^2 |f|_{2,\Omega}$$

completing the proof.

To obtain an estimate for the 7-point numerical quadrature error, we have the following Theorem.

Theorem 3.4.2. If $f \in H^4(\Omega)$ and $\tilde{F}_\alpha$ is the 7-point numerical quadrature of F_α, then

$$\Big(\sum_{\alpha\in\mathring{\Gamma}_h} |E(f,X_\alpha)|^2\Big)^{\frac{1}{2}} \le 0.07208(\mu_\Omega(T))^{\frac{1}{2}} h^4 |f|^2_{4,\Omega} \qquad (3.4.3)$$

To prove the theorem, we need the following auxiliary lemmas.

Lemma 3.4.3. If $f \in H^4(\Omega)$ and $P(\xi_0)$ is a real valued function of ξ_0 defined on each of the triangular element $T_j = X_0X_jX_{j+1}$, then,

$$\sum_{j=1}^{6}\Big[\int_{X_0}^{X_j} P(\xi_0)D_{200}f\, dX - \int_{X_{j+1}}^{X_0} P(\xi_0)D_{100}f\, dX\Big]$$

$$= \sum_{j=1}^{6}\int_{T_j} \frac{1}{2}P(\xi_0)(D_{0000}f - \frac{1}{2}D_{0012}f)\, d\mu_{T_j} \qquad (3.4.4)$$

Proof :

$$\sum_{j=1}^{6}\Big[\int_{X_0}^{X_j} P(\xi_0)D_{200}f\, dX - \int_{X_{j+1}}^{X_0} P(\xi_0)D_{100}f\, dX\Big]$$

$$= \sum_{j=1}^{6}\Big[\int_{X_0}^{X_j} P(\xi_0)(-D_{000}f - D_{100}f)\, dX - \int_{X_{j+1}}^{X_0} P(\xi_0)(-D_{000}f - D_{200}f)\, dX\Big]$$

$$= \sum_{j=1}^{6} [\int_{X_{j+1}}^{X_0} P(\xi_0)D_{000}f\ dX - \int_{X_0}^{X_j} P(\xi_0)D_{000}f\ dX +$$

$$\int_{X_{j+1}}^{X_0} P(\xi_0)D_{200}f\ dX - \int_{X_0}^{X_j} P(\xi_0)D_{100}f\ dX]$$

$$= \sum_{j=1}^{6} [\int_{T_j} \frac{1}{2}P(\xi_0)D_{0000}f\ d\mu_{T_j} + \int_{X_{j+1}}^{X} P(\xi_0)D_{200}f\ dX - \int_{X_0}^{X_j} P(\xi_0)D_{100}f\ dX]$$

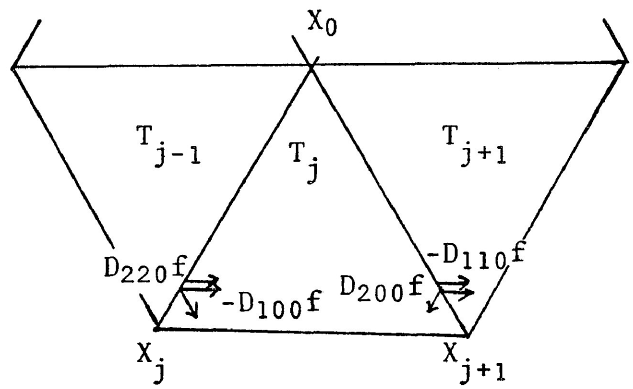


Fig. 3.4.1

As shown in Fig. 3.4.1, the derivative $-D_{100}f$ along X_jX_0 w.r.t. T_j is the same as $D_{220}f$ w.r.t. T_{j-1}, and $D_{200}f$ along $X_{j+1}X_0$ w.r.t. T_j is the same as $-D_{110}f$ w.r.t. T_{j+1}. Thus, these derivatives can be divided into two equal parts, half of them will be added to the line integral of the adjacent triangle. It follows that

$$\sum_{j=1}^{6} [\int_{X_0}^{X_j} P(\xi_0) D_{200} f \, dX - \int_{X_{j+1}}^{X_0} P(\xi_0) D_{100} f \, dX]$$

$$= \sum_{j=1}^{6} [\int_{T_j} \frac{1}{2} P(\xi_0) D_{0000} f \, d\mu_{T_j} + \int_{X_{j+1}}^{X_0} \frac{1}{2} P(\xi_0)(D_{200} f + D_{220} f) \, dX -$$

$$\int_{X_0}^{X_j} \frac{1}{2} P(\xi_0)(D_{100} f + D_{110} f) \, dX]$$

$$= \sum_{j=1}^{6} [\int_{T_j} \frac{1}{2} P(\xi_0) D_{0000} f \, d\mu_{T_j} - \int_{X_{j+1}}^{X_0} \frac{1}{2} P(\xi_0) D_{210} f \, dX + \int_{X_0}^{X_j} \frac{1}{2} P(\xi_0) D_{120} f \, dX]$$

$$= \sum_{j=1}^{6} \int_{T_j} [\frac{1}{2} P(\xi_0) D_{0000} f - \frac{1}{4} P(\xi_0) D_{0012} f] \, d\mu_{T_j}$$

completing the proof.

Lemma 3.4.4. If $f \in H^4(\Omega)$, then the error functional of the 7-point numerical quadrature has a representation of the form

$$E(f, X_\alpha) = \int_\Omega f \phi_\alpha d\mu_\Omega - \mu_\Omega(T)[\frac{3}{2} f(X_0) + \frac{1}{12} \sum_{j=1}^{6} f(X_j)]$$

$$= \frac{1}{24} \mu_\Omega(T) \sum_{j=1}^{6} \int_{T_j} \{[(\frac{1}{2} + \xi_0 - 8\xi_0^2 + 5\xi_0^3 + \xi_0 \xi_1 \xi_2)\xi_1 \xi_2 - \frac{\xi_0}{2} - \frac{\xi_0^2}{4} +$$

$$3\xi_0^3 - \frac{13}{4}\xi_0^4 + \xi_0^5] D_{0000} f + \frac{1}{2}(\frac{\xi_0}{2} + \frac{\xi_0^2}{4} - 3\xi_0^3 + \frac{13}{4}\xi_0^4 - \xi_0^5) D_{0012} f\} \, d\mu_{T_j}$$

(3.4.5)

Proof: We shall only give a brief proof for this lemma.

$$E(f,X_\alpha) = [\int_\Omega f\phi_\alpha d\mu_\Omega - 2h^2 f(X_\alpha)] + \frac{\mu_\Omega(T)}{12}\sum_{j=1}^{6}[f(X_0)-f(X_j)]$$

$$= [\int_\Omega f\phi_\alpha d\mu_\Omega - 2h^2 f(X_\alpha)] + \frac{1}{12}\mu_\Omega(T)\sum_{j=1}^{6}\int_{X_j}^{X_0} D_{X_jX_0} f\, dX$$

It follows from Lemma 3.4.2 and Lemma 3.4.1 that

$$E(f,X_\alpha) = \sum_{j=1}^{6}\frac{\mu_\Omega(T)}{24}\int_{T_j}(-1+12\xi_0^2-8\xi_0^3-12\xi_0\xi_1\xi_2)D_{00}f\, d\mu_{T_j}$$

It is not hard to get into the following step:

$$E(f,X_\alpha) = \frac{\mu_\Omega(T)}{24}\sum_{j=1}^{6}[\int_{X_{j+1}}^{X_0}(-1-2\xi_0+16\xi_0^2-10\xi_0^3)\xi_2 D_{00}f\, dX +$$

$$\int_{X_0}^{X_j}(-1-2\xi_0+16\xi_0^2-10\xi_0^3)\xi_1 D_{00}f\, dX -$$

$$\int_{T_j}\frac{1}{2}(-1-2\xi_0+16\xi_0^2-10\xi_0^3)(\xi_2-\xi_1)D_{000}f\, d\mu_{T_j} -$$

$$\int_{T_j} D_0(\xi_0\xi_1^2\xi_2^2)D_{000}f\, d\mu_{T_j}]$$

after further evaluation, we have

$$E(f,X_\alpha) = \frac{\mu_\Omega(T)}{24}\sum_{j=1}^{6}\Big[\int_{X_0}^{X_j}(-\xi_0-\frac{\xi_0^2}{2}+6\xi_0^3-\frac{13}{2}\xi_0^4+2\xi_0^5)D_{200}f\ dX\ -$$

$$\int_{X_{j+1}}^{X_0}(-\xi_0-\frac{\xi_0^2}{2}+6\xi_0^3-\frac{13}{2}\xi_0^4+2\xi_0^5)D_{100}f\ dX\ +$$

$$\int_{T_j}(\frac{1}{2}+\xi_0-8\xi_0^2+5\xi_0^3+\xi_0\xi_1\xi_2)\xi_1\xi_2 D_{0000}f\ d\mu_{T_j}$$

Applying the Lemma 3.4.3, the result of Lemma 3.4.4 follows.

Proof of Theorem 3.4.2:

Applying the triangle inequality and the Cauchy-Schwarz inequality to equation (3.4.5), we get

$$|E(f,X_\alpha)| \le \frac{\mu_\Omega(T)}{24}\sum_{j=1}^{6}(k_0\|D_{0000}\,f\|_{L^2(T_j)}+k_1\|D_{0012}f\|_{L^2(T_j)})$$

where $k_0 = 0.08495$ and $k_1 = 0.04297$

Applying the Cauchy-Schwarz inequality to the above inequality, we get

$$|E(f,X_\alpha)| \le \frac{\mu_\Omega(T)}{24}(6k_0^2+6k_1^2)^{\frac{1}{2}}\{\sum_{j=1}^{6}(\|D_{0000}f\|^2_{L^2(T_j)}+\|D_{0012}f\|^2_{L^2(T_j)})\}^{\frac{1}{2}}$$

It follows that

$$(\sum_{\alpha\in\mathring{\Gamma}_h}|E(f,X_\alpha)|^2)^{\frac{1}{2}} \le 0.07208(\mu_\Omega(T))^{\frac{1}{2}}\{\sum_{\alpha\in\mathring{\Gamma}_h}\mu_\Omega(T)\sum_{j=1}^{6}(\|D_{0000}f\|^2_{L^2(T_j)}+ \|D_{0012}f\|^2_{L^2(T_j)}\}^{\frac{1}{2}} \qquad (3.4.6)$$

for each $T \in \tau^h$, since the terms $\|D_{0000}f\|^2_{L^2(T_j)}$ and $\|D_{0012}f\|^2_{L^2(T_j)}$ appear in the right hand side of the inequality (3.4.6) at most once, thus, we have

$$\Big(\sum_{\alpha\in\mathring{\Gamma}_h} |E(f,X_\alpha|^2\Big)^{\frac{1}{2}} \le 0.07208(\mu_\Omega(T))^{\frac{1}{2}}\Big(\sum_{T\in\tau^h} \mu_\Omega(T) \sum_{|\beta|=4} \|D^\beta f\|^2_{L^2(T)}\Big)^{\frac{1}{2}}$$

$$= 0.07208(\mu_\Omega(T))^{\frac{1}{2}} h^4 |f|_{4,\Omega}$$

completing the proof.

We know from Section 2.5 that the Ritz-Galerkin solution to the linear operator $Lu = -\nabla\cdot(p\nabla u)+qu = f$ turns out to solve the following system of linear equations:

$$\int_\Omega (p\nabla u^h\cdot\nabla\phi_\alpha + qu^h\phi_\alpha)\, d\mu_\Omega = \int_\Omega f\phi_\alpha d\mu_\Omega , \qquad \alpha \in \mathring{\Gamma}_h$$

or it can be written as

$$a(u^h,\phi_\alpha) = F_\alpha = (f,\phi_\alpha) \tag{3.4.7}$$

If the integral F_α is approximated by a numerical quadrature $\tilde{F}_\alpha$, then we are solving

$$a(\tilde{u}^h,\phi_\alpha) = \tilde{F}_\alpha \qquad \alpha \in \mathring{\Gamma}_h \tag{3.4.8}$$

where $\tilde{u}^h = \sum_{\alpha\in\mathring{\Gamma}} \tilde{\lambda}_\alpha \phi_\alpha$ is a solution to the linear system (3.4.8).

From (3.4.7) and (3.4.8) we get

$$a(u^h-\tilde{u}^h,\phi_\alpha) = F_\alpha-\tilde{F}_\alpha = E(f,X_\alpha)$$

It follows that

$$a(u^h-\tilde{u}^h,u^h-\tilde{u}^h) = \sum_{\alpha\in\mathring{\Gamma}_h}(\lambda_\alpha-\tilde{\lambda}_\alpha)E(f,X_\alpha)$$

this reduces to

$$\| u^h-\tilde{u}^h\|_a^2 \le \sum_{\alpha\in\mathring{\Gamma}_h} |\lambda_\alpha-\tilde{\lambda}_\alpha| \; |E(f,X_\alpha)| \tag{3.4.9}$$

By applying the Cauchy-Schwarz inequality to the equation (3.4.9), we get

$$\| u^h-\tilde{u}^h\|_a^2 \le \Big(\sum_{\alpha\in\mathring{\Gamma}_h} (\lambda_\alpha-\tilde{\lambda}_\alpha)^2\Big)^{\frac{1}{2}}\Big(\sum_{\alpha\in\mathring{\Gamma}_h} |E(f,X_\alpha)|^2\Big)^{\frac{1}{2}} \tag{3.4.10}$$

To obtain an upper bound for $\Big(\sum_{\alpha\in\mathring{\Gamma}_h} (\lambda_\alpha-\tilde{\lambda}_\alpha)^2\Big)^{\frac{1}{2}}$ in terms of the L^2 norm $\| u^h-\tilde{u}^h \|_{L^2(\Omega)}$, we need the following Lemma.

<u>Lemma 3.4.5.</u> Let $u(X) = \sum_{\alpha\in\Gamma_h} \lambda_\alpha\phi_\alpha(X)$ and vanishes on $\partial\Omega$. Then

$$\| u \|_{L^2(\Omega)}^2 \ge \frac{1}{2}\mu_\Omega(T) \sum_{\alpha\in\mathring{\Gamma}_h} \lambda_\alpha^2$$

<u>Proof</u> : $\|u\|^2_{L^2(\Omega)} = \int_\Omega u^2 \, d\mu_\Omega$

$$= \mu_\Omega(T) \sum_{T\in\tau^h} \int_T (\lambda_\alpha\phi_\alpha+\lambda_\beta\phi_\beta+\lambda_\gamma\phi_\gamma)^2 \, d\mu_T$$

$$= \frac{1}{6}\mu_\Omega(T) \sum_{T\in\tau^h} (\lambda_\alpha^2+\lambda_\beta^2+\lambda_\gamma^2+\lambda_\beta\lambda_\gamma+\lambda_\gamma\lambda_\alpha+\lambda_\alpha\lambda_\beta)$$

where λ_α, λ_β, λ_γ are the values of u at the three vertices of $T = X_\alpha X_\beta X_\gamma$.

Since $\lambda_\alpha^2 + \lambda_\beta^2 + \lambda_\gamma^2 + 2(\lambda_\beta\lambda_\gamma+\lambda_\gamma\lambda_\alpha+\lambda_\alpha\lambda_\beta) \geq 0$ for all $\lambda_\alpha, \lambda_\beta, \lambda_\gamma \in R$, we have

$$\sum_{T\in\tau^h} (\lambda_\alpha^2+\lambda_\beta^2+\lambda_\gamma^2+\lambda_\beta\lambda_\gamma+\lambda_\gamma\lambda_\alpha+\lambda_\alpha\lambda_\beta) \geq \sum_{T\in\tau^h} \frac{1}{2}(\lambda_\alpha^2+\lambda_\beta^2+\lambda_\gamma^2) \qquad (3.4.11)$$

Since $\lambda_\alpha = 0$ for all $X_\alpha \in \partial\Omega$, and for each $\alpha \in \mathring{\Gamma}_h$, there are six triangles T in τ^h with the common vertex X_α, thus the right hand side of the inequality (3.4.11) can be written as $3 \sum_{\alpha\in\mathring{\Gamma}_h} \lambda_\alpha^2$. It follows that

$$\|u\|^2_{L^2(\Omega)} \geq \frac{1}{2}\mu_\Omega(T) \sum_{\alpha\in\mathring{\Gamma}_h} \lambda_\alpha^2 ,$$

completing the proof.

Since $u^h-\tilde{u}^h$ is a piecewise linear function on Ω and vanishes on the boundary $\partial\Omega$, we can apply Lemma 3.4.5 to the inequality (3.4.10) to get

$$\|u^h-\tilde{u}^h\|_a^2 \le \left(\frac{2}{\mu_\Omega(T)}\right)^{\frac{1}{2}} \|u^h-\tilde{u}^h\|_{L^2(\Omega)} \left(\sum_{\alpha\in\tilde{\Gamma}_h} |E(f,X_\alpha)|^2\right)^{\frac{1}{2}}$$

From Lemma 2.3.2 we have

$$\|u^h-\tilde{u}^h\|_a \ge \sigma\|u^h-\tilde{u}^h\|_{1,\Omega} \ge \sigma\|u^h-\tilde{u}^h\|_{L^2(\Omega)}$$

It follows that

$$\|u^h-\tilde{u}^h\|_a \le \frac{\sqrt{2}}{\sigma(\mu_\Omega(T))^{\frac{1}{2}}} \left(\sum_{\alpha\in\tilde{\Gamma}_h} |E(f,X_\alpha)|^2\right)^{\frac{1}{2}}$$

If the 1-point numerical quadrature is used, from Theorem 3.4.1 we have

$$\|u^h-\tilde{u}^h\|_a \le \left(\frac{23}{280}\right)^{\frac{1}{2}} \frac{1}{\sigma} h^2 |f|_{2,\Omega} ,$$

and if the 7-point numerical quadrature is used, from Theorem 3.4.2 we have

$$\|u^h-\tilde{u}^h\|_a \le \frac{0.1019}{\sigma} h^4 |f|_{4,\Omega}$$

If u is the exact solution to $Lu = f$, from the triangle inequality, we get

$$||u-\tilde{u}^h||_a \leq ||u-u^h||_a + ||u^h-\tilde{u}^h||_a$$

From Theorem 3.3.2 we know that the energy norm $||u-u^h||_a$ has an order of accuracy $O(h)$, whereas the energy norm $||u^h-\tilde{u}^h||_a$ has an order of accuracy $O(h^2)$ and $O(h^4)$ for the 1-point and 7-point numerical quadrature respectively. Thus, both the numerical quadratures are consistent [V2] in the energy norm, that is, the solution still has an order of accuracy $O(h)$ in the energy norm for the 1-point and 7-point numerical quadrature.

CHAPTER 4

SOLUTION OF THE DISCRETE LINEAR EQUATIONS

4.1 INTRODUCTION

It is well known that discrete two dimensional boundary value problems become very hard to solve by the usual iterative algorithms as the number n of data points become large. P.O. Frederickson has introduced an algorithm FAPIN [F4] to solve this type of problem. In particular, the algorithm FAPIN solves the Ritz-Galerkin approximation in O(n) operations and O(n) storages.

In this chapter, we lean heavily on the first few sections of Frederickson [F4] and many of our results come from this source.

The algorithm FAPIN requires an approximate 1-local inverse C. This approximate inverse can be constructed by the TRq or LSq method introduced by Benson [B3]. The TRq method is generalized to the weighted truncation (WTq) method by multiplying a weight W to $CA-I$.

We then introduce a new technique for the construction of an optimal ε-approximate inverse to A, which we refer to as the *interpolation method*, (INq). Numerical results with each approximate inversion technique considered are presented, serving as a basis of comparison of different constructive methods.

We end this chapter by presenting some numerical examples for the solving of the Poisson equation in a triangular domain with homogeneous and inhomogeneous boundary conditions, and in one of these, the differential operator is singular.

4.2 APPROXIMATE INVERSION

Let $\|\cdot\|_x$ and $\|\cdot\|_y$ be the norms of the Banach spaces X and Y respectively, and let A be a bounded linear operator mapping X into Y. For a given y in the range of A, we are interested in constructing a numerical solution $x \in X$ s.t.

$$Ax = y \tag{4.2.1}$$

We recall two definitions from Frederickson [F4]:

Definition 4.2.1. Given $0 < \varepsilon < 1$, then an element $x \in X$ is called an *ε-approximate solution* to (4.2.1) if

$$\|y-Ax\|_y \le \varepsilon\|y\|_y \tag{4.2.2}$$

Definition 4.2.2. For $0 < \varepsilon < 1$, a linear operator $C : Y \to X$ is called an *ε-approximate inverse* to A if

$$\|Ax-ACAx\|_y \le \varepsilon\|Ax\|_y \quad \text{for all } x \in X \tag{4.2.3}$$

If A is nonsingular, then (4.2.3) is equivalent to the inequality

$$\|I-AC\| \le \varepsilon \tag{4.2.4}$$

which is known ([F7], [V1]) to be a sufficient condition for the convergence of the iterative process

$$\begin{cases} r_k = y - Ax_k \\ x_{k+1} = x_k + Cr_k \end{cases} \tag{4.2.5}$$

to a solution to (4.2.1) for any initial approximation x_0 and any y in the range of A.

If A is singular, Frederickson [F7] has shown that the iterative process (4.2.5) still works, provided only that (4.2.1) has a solution.

Theorem 4.2.1. [F7] If C is a nonsingular ε-approximate inverse to A, then the following are equivalent:

(a) There exists an $x_0 \in X$, such that the iteration procedure (4.2.5) converges

(b) Equation (4.2.1) has a solution

(c) For any starting vector $x_0 \in X$, the sequence $<x_k>$ of (4.2.5) converges to a solution to (4.2.1), and the map : $x_0 \to x$ is affine and onto the set of all solutions to (4.2.5).

Proof : (a) $\Longrightarrow$ (b)

Let x be an element of X s.t.

$x_k \to x$

From (4.2.5) we have

$$Cr_k = x_{k+1} - x_k \to 0$$

$$r_k = y - Ax_k \to y - Ax$$

from which it follows that

$$C(y-Ax) = 0$$

Since C is nonsingular, we have

$$y = Ax$$

Now we want to prove (b) $\Longrightarrow$ (c)

Let $x^* \in X$ s.t. $Ax^* = y$

From (4.2.5) we have

$$r_{k+1} = y - Ax_{k+1}$$

$$= y - A(x_k+Cr_k)$$

$$= (y-Ax_k) - AC(y-Ax_k)$$

$$= A(x^*-x_k) - ACA(x^*-x_k)$$

Since C is an ε-approximate inverse to A, we have

$$\|r_{k+1}\|_y \leq \varepsilon\|A(x^* - x_k)\|_y = \varepsilon\|r_k\|_y$$

It follows that

$$\|r_k\|_y \leq \varepsilon^k\|r_0\|_y \tag{4.2.6}$$

From (4.2.5), we have

$$\|x_m - x_n\|_x = \|Cr_{m-1} + Cr_{m-2} + \cdots + Cr_n\|_x \quad \forall \quad m > n$$

$$\leq \|C\|(\|r_{m-1}\|_y + \|r_{m-2}\|_y + \cdots + \|r_n\|_y)$$

$$\leq \|C\| \ \|r_0\|_y (\varepsilon^{m-1}+\varepsilon^{m-2}+ \cdots +\varepsilon^n)$$

$$< \|C\| \ \|r_0\|_y \ \varepsilon^n/(1-\varepsilon) \to 0 \quad \text{as} \quad n \to \infty \quad \text{which}$$

implies $\langle x_k \rangle$ is a Cauchy sequence in X and hence converges to a point $x \in X$.

Thus $Ax_k \to Ax$

From (4.2.6), we have $r_k \to 0$, hence

$$y - Ax_k \to 0$$

or $$Ax_k \to y$$

It follows that $Ax = y$

To prove that the map described by (4.2.5) is affine, let $x_{1,k}$ and $x_{2,k}$ be any two elements of X and

$$\lambda_1 + \lambda_2 = 1$$

then from

$$x_k = \lambda_1 x_{1,k} + \lambda_2 x_{2,k} \quad \text{follows}$$

$$x_{k+1} = x_k + C(y-Ax_k)$$

$$= \lambda_1 x_{1,k} + \lambda_2 x_{2,k} + \lambda_1 Cy + \lambda_2 Cy - \lambda_1 CAx_{1,k} - \lambda_2 CAx_{2,k}$$

$$= \lambda_1[x_{1,k} + C(y-Ax_{1,k})] + \lambda_2[x_{2,k} + C(y-Ax_{2,k})]$$

$$= \lambda_1 x_{1,k+1} + \lambda_2 x_{2,k+1}$$

Thus the map $x_0 \to x$ described by (4.2.5) is an affine map. To prove that it is onto the range of A is easy, if $Ax = y$ we simply choose $x_0 = x$.

The implication from (c) to (a) is trival, completing the proof.

Define by $\rho_m = \dfrac{\| r_m \|_y}{\| r_{m-1} \|_y}$ the *reduction factor* [V1] at iteration m, if $\|r_{m-1}\| \neq 0$, where r_m is the residual vector defined in (4.2.5).

If the largest eigenvalue λ in modulus of the linear operator $I-AC$ is dominant, and if r_0 is not orthogonal to the eigenvector V corresponding to λ, then the limit of ρ_m exists and [B5, p. 269]

$$\lim_{m\to\infty} \rho_m = \rho(I-AC)$$

Thus, the spectral radius $\rho(I-AC)$ serves as a basis of comparison of how well the operator C approximates the inverse of

A in an iterative algorithm.

In terms of actual computations, the spectral radius ρ of the operator I-AC can be estimated from the computation of the reduction factor ρ_m in an iterative algorithm to the solution of the equation

$$Ax = 0$$

with a random initial vector x.

If we order the eigenvalues of the operator I-AC so that

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n|$$

Then the rate at which the sequence ρ_m converges depends on the dominance ratio : [V1]

$$\delta = \frac{|\lambda_2|}{|\lambda_1|} < 1$$

the convergence of the estimate ρ_m of ρ is slow when δ is close to 1. However, the convergence of the sequence ρ_m can be accelerated by the application of a non-linear sequence-to-sequence transformations proposed by D. Shanks [S3]. A Fortran program to perform this transformation is given in Appendix B.

4.3 LOCAL OPERATORS

For the purpose of solving the system of linear equations produced by the Ritz-Galerkin solution to the linear operator

$Lu = f$ on a bounded polygonal domain Ω, we restrict our attention to finite dimensional linear spaces X and Y.

Denote by X_{Γ_h} the space of real valued functions on the integer lattice Γ_h defined in section 1.2, and let Y_{Γ_h} be a subspace of X. We say that the linear operator $A : X_{\Gamma_h} \to Y_{\Gamma_h}$ is a q-*local operator* for some integer q if the value of Ax at a point $\alpha \in \Gamma_h$ depends only on the values of x in a q-neighbourhood of α ; [F4], more precisely, if

$$[(Ax)_\alpha \neq 0] \Rightarrow [\exists\ \beta \in \Gamma_h,\ \ |\alpha-\beta| \leq q, \text{ and } x_\beta \neq 0],$$

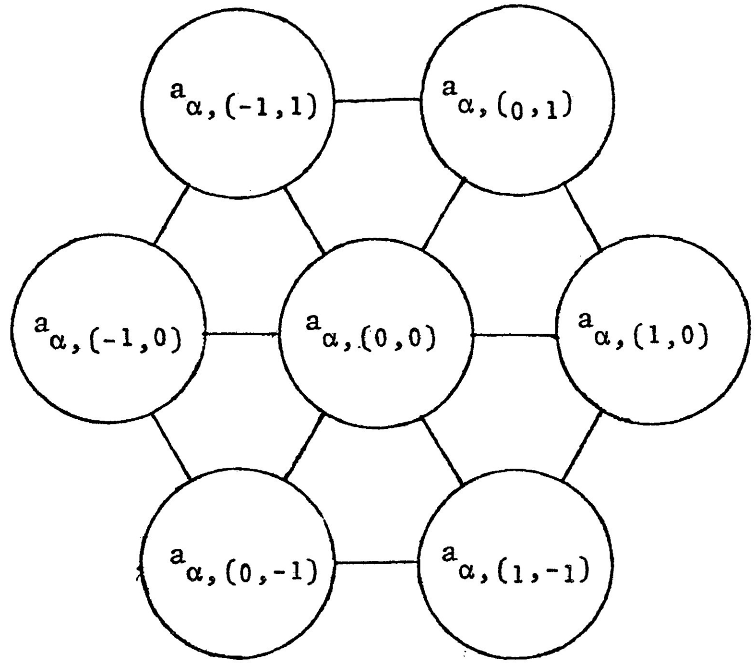
where $|\cdot|$ is the hexagonal norm defined in Section 1.2.

Thus, for any q-local operator $A : X_{\Gamma_h} \to Y_{\Gamma_h}$, there are elements $a_{\alpha,\beta}$ s.t. for any point $\alpha \in \Gamma_h$

$$(Ax)_\alpha = \sum_{|\beta| \leq q} a_{\alpha,\beta} x_{\alpha+\beta} \tag{4.3.1}$$

In particular, if A is a 1-local operator, then at each point $\alpha \in \mathring{\Gamma}_h$, expressed diagrammatically, A has a representation of the form

A :



Denote by n the number of points in Γ_h, then the implementation of (4.3.1) allows storage of A in $7n$ locations and evaluation of Ax in $7n$ multiplications.

Let A be a q_1-local operator and C be a q_2-local operator defined on the linear space X_{Γ_h}. We seek the linear operator B such that for any $x \in X_{\Gamma_h}$

$$Bx = C(Ax)$$

In terms of the representation (4.3.1), B can be expressed as

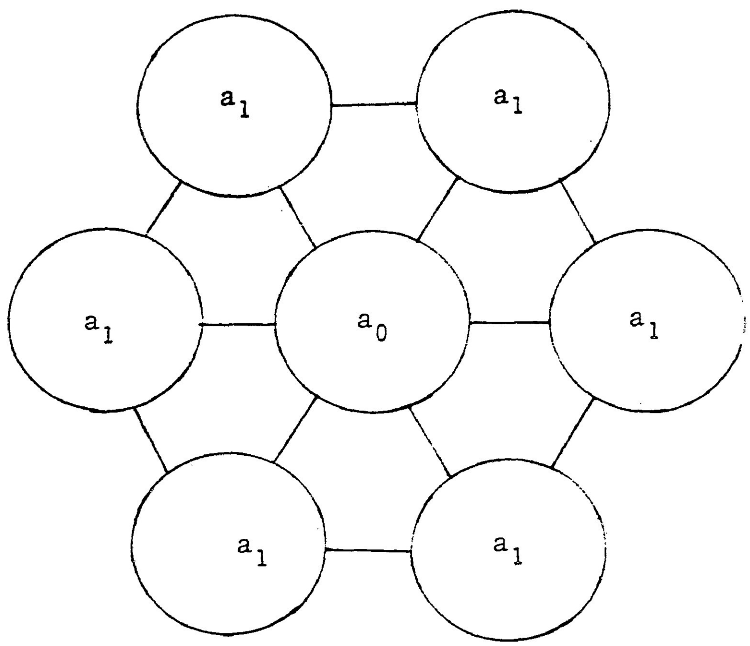
$$(Bx)_\alpha = \sum_{\substack{\beta,\gamma \\ |\beta|\le q_1, |\gamma|\le q_2}} c_{\alpha,\beta} a_{\alpha+\beta,\gamma} x_{\alpha+\beta+\gamma} \qquad (4.3.2)$$

The sum extends over only those α and β for which $\alpha + \beta \in \Gamma_h$.

As we can see from (4.3.2), β is a (q_1+q_2)-local operator.

In particular, if A is a constant coefficient 1-local operator with a representation of the form

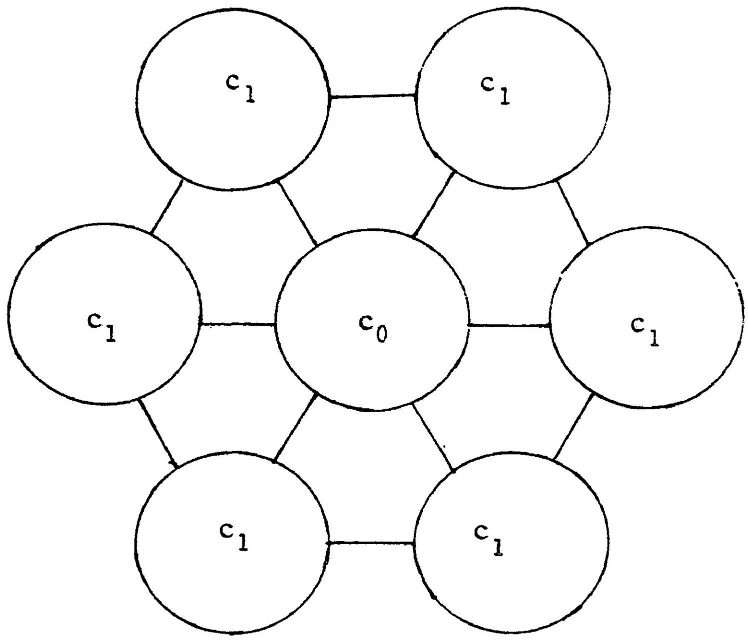
A :



(4.3.3)

and C is also a constant coefficient 1-local operator with a representation of the form

C :

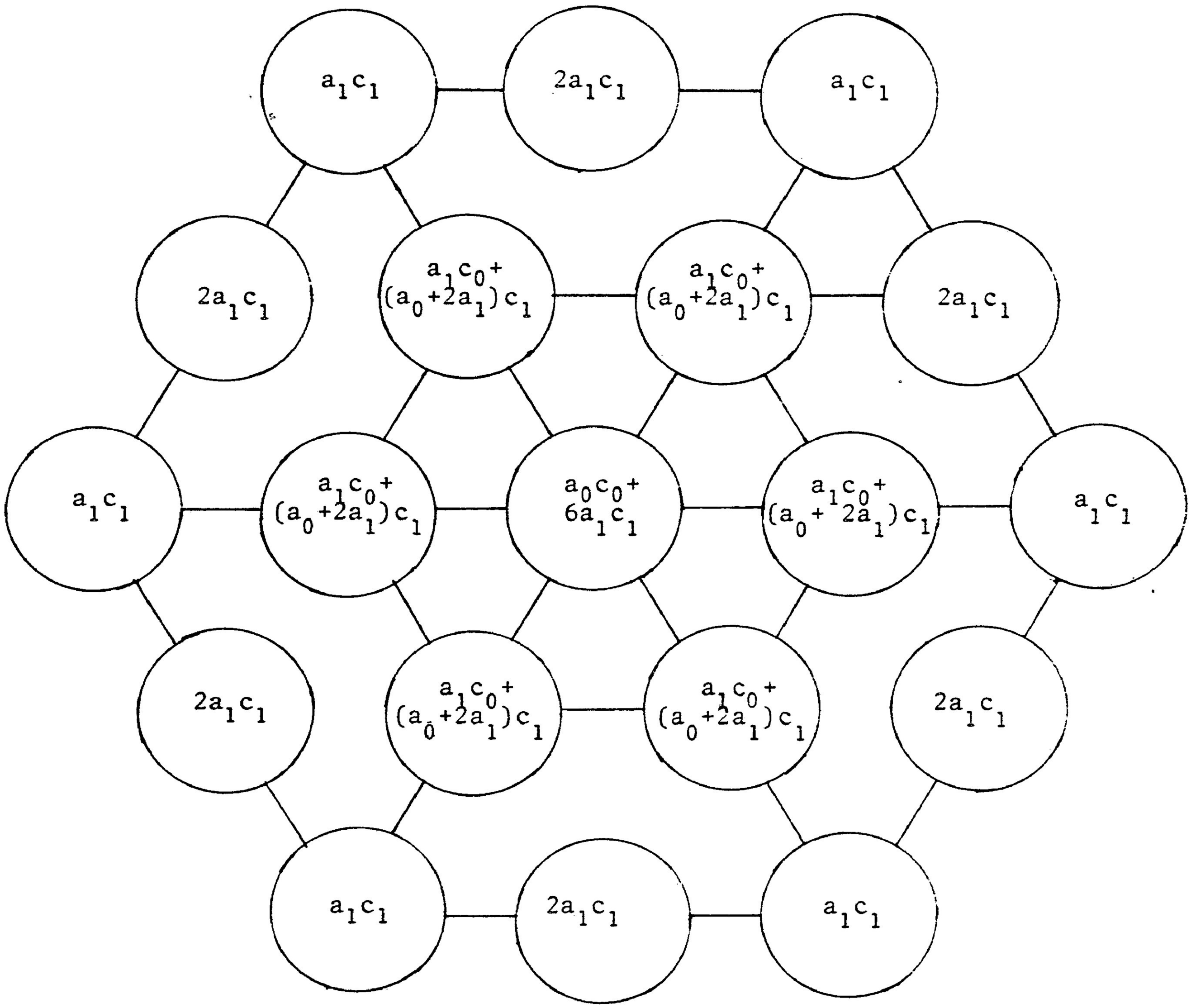


(4.3.4)

then the composition of CA is a 2-local operator. The graph of B = CA is shown in (4.3.5).

It is easy to see that if A and C are constant coefficient local operators, then the composition commutes, i.e. AC = CA.

In this case, AC can be written as a convolution operator.



a1c1
2a1c1
a1c1
2a1c1
a1c0+
(a0+2a1)c1
a1c0+
(a0+2a1)c1
2a1c1
a1c1
a1c0+
(a0+2a1)c1
a0c0+
6a1c1
a1c0+
(a0+2a1)c1
a1c1
2a1c1
a1c0+
(a0+2a1)c1
a1c0+
(a0+2a1)c1
2a1c1
a1c1
2a1c1
a1c1

(4.3.5)

4.4 BEST APPROXIMATION

For every triangulation τ^h of Ω there is a least integer ℓ such that $|\alpha|\leq 2^{\ell-1}$ for every $\alpha \in \Gamma_h$, we write Γ^ℓ for Γ_h and define, using the recurrence

$$\Gamma^{k-1} = \{\alpha : \exists \beta,\ |\beta|\leq 1,\ 2\alpha+\beta \in \Gamma^k\} \tag{4.4.1}$$

the sets Γ^k for $1\leq k\leq \ell$

We observe that $|\alpha|\leq 2^{k-1}$ if $\alpha \in \Gamma^k$, and in particular, Γ^1 has at most 7 points.

Denote by X^k the linear space of real valued functions defined on Γ^k, and define the sequence of *interpolation operators* $Q^k : X^{k-1} \to X^k$ through

$$x^k_\beta = (Q^k x^{k-1})_\beta = \sum_{\alpha\in\Gamma^{k-1}} \phi^k_\alpha(\beta) x^{k-1}_\alpha \tag{4.4.2}$$

where

$$\phi^k_\alpha(\beta) = \begin{cases} \frac{1}{2} & \text{if } |2\alpha-\beta| = 1 \\ 1 & \text{if } \beta = 2\alpha \\ 0 & \text{otherwise} \end{cases}$$

The set $\{\phi^k_\alpha\}_{\alpha\in\Gamma^{k-1}}$ form a basis for the space

$U^k = Q^k(X^{k-1})$.

Define the sequence of *projection operators*

$$P^k : X^k \to X^{k-1} \quad \text{by}$$

$$r_\alpha^{k-1} = (P^k r^k)_\alpha = \sum_{\beta\in\Gamma^k} \phi_\alpha^k(\beta) r_\beta^k \tag{4.4.3}$$

Beginning with $A^\ell = A$ and $Y^\ell = Y$, we define the sequence of operators $A^k : X^k \to Y^k$ by

$$A^{k-1} = P^k A^k Q^k : X^{k-1} \to Y^{k-1} \tag{4.4.4}$$

Then in terms of the representation (4.3.1), A^{k-1} can be represented as

$$(A^{k-1}x^{k-1})_\alpha = \sum_{|\gamma|\le 1} \phi_\alpha^k(2\alpha+\gamma) \sum_{|\sigma|\le 1} a^k_{2\alpha+\gamma,\sigma} \sum_{\substack{\beta \\ |\gamma+\sigma-2\beta|\le 1}} \phi^k_{\alpha+\beta}(2\alpha+\gamma+\sigma) x^{k-1}_{\alpha+\beta}$$

$$= \sum_{\substack{\beta,\gamma,\sigma \\ |\beta|\le 1, |\gamma|\le 1, |\sigma|\le 1 \\ |\gamma+\sigma-2\beta|\le 1}} \phi_\alpha^k(2\alpha+\gamma) a^k_{2\alpha+\gamma,\sigma} \phi^k_{\alpha+\beta}(2\alpha+\gamma+\sigma) x^{k-1}_{\alpha+\beta}$$

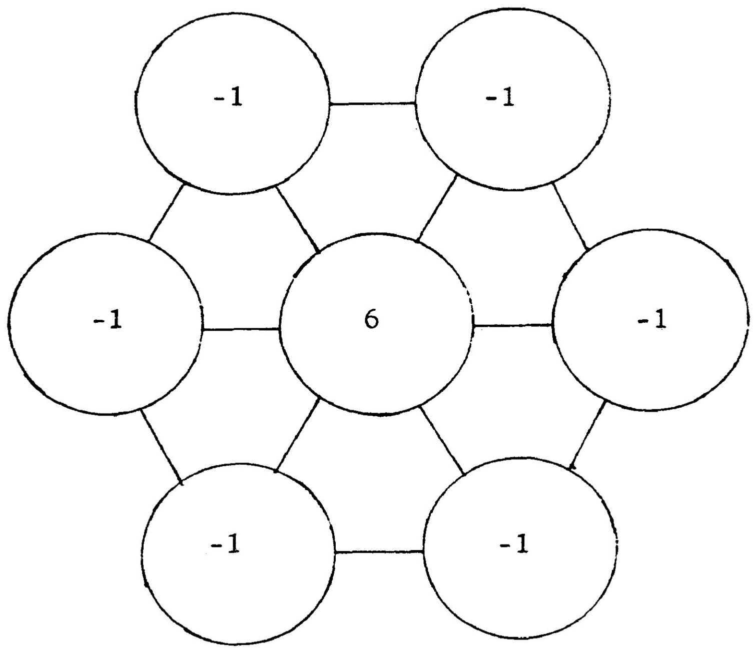
or

$$a^{k-1}_{\alpha,\beta} = \sum_{\substack{\gamma,\sigma \\ |\gamma|\le 1, |\sigma|\le 1 \\ |\gamma+\sigma-2\beta|\le 1}} \phi_\alpha^k(2\alpha+\gamma) a^k_{2\alpha+\gamma,\sigma} \phi^k_{\alpha+\beta}(2\alpha+\gamma+\sigma) \tag{4.4.5}$$

In particular, if A is the 1-local discrete Lapacian operator derived in (2.5.6), it is easy to verify that A is invariant under the collection.

Thus, we have

$A^k = A$

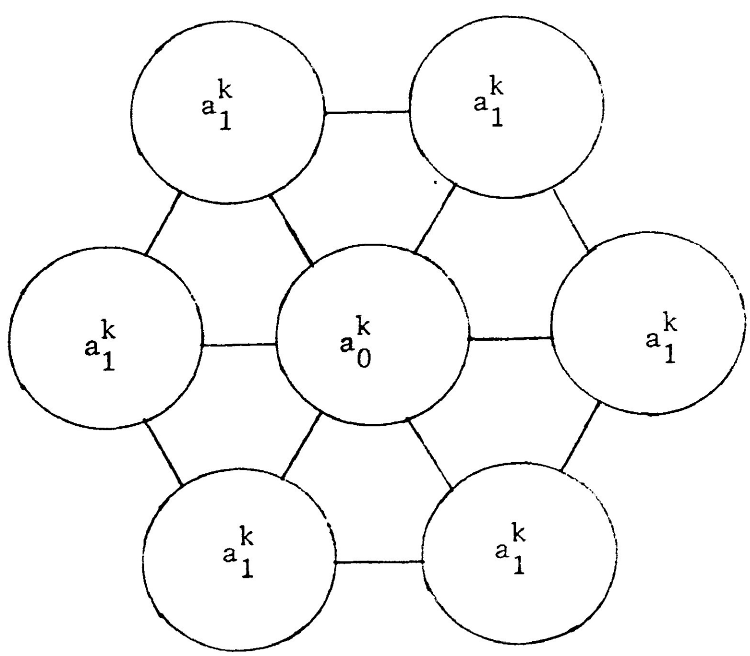


(4.4.6)

for all $1 \le k \le \ell$

If A^k is a constant coefficient operator of the form

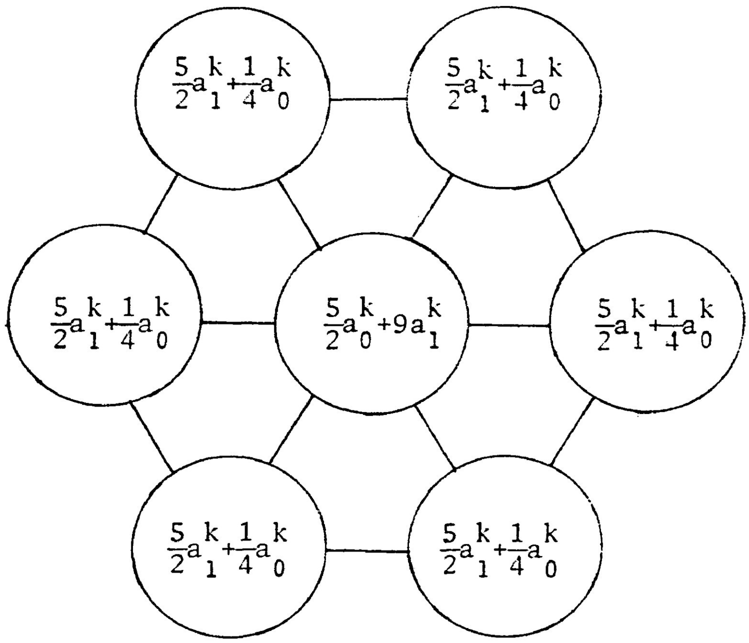
A^k :



(4.4.7)

then we have

A^{k-1}



(4.4.8)

It follows that $A^{k-1} = \lambda A^k$ for some constant λ iff

$$\begin{cases} \frac{5}{2}a_0^k + 9a_1^k = \lambda a_0^k \\ \frac{5}{2}a_1^k + \frac{1}{4}a_0^k = \lambda a_1^k \end{cases}$$

iff $\lambda = 4$ or 1

The two constant coefficient a_0^ℓ and a_1^ℓ of A^ℓ are related by

$$\begin{cases} a_0^\ell = -6a_1^\ell & \text{if } \lambda = 1 \\ a_0^\ell = 6a_1^\ell & \text{if } \lambda = 4 \end{cases}$$

Given an element r^k in the range of A^k, we are asked to find an element $x^k \in X^k$ s.t.

$$A^k x^k = r^k \tag{4.4.9}$$

Let r^{k-1} be the image of P^k at r^k, if x^{k-1} is the the solution of the equation

$$A^{k-1}x^{k-1} = r^{k-1}$$

We are interested to know how close the solution x^{k-1} is to x^k? This question is answered by the following Theorem [F4]:

Theorem 4.4.1. If A is symmetric and positive definite, then the operator A^{k-1} defined by (4.4.4) is the Ritz-Galerkin best approximation to A^k in the subspace $U^k = Q^k(X^{k-1})$ of X^k.

Proof: We define the quadratic functional related to (4.4.9) by

$$F(x^k) = \langle Ax^k - 2r^k, x^k \rangle$$

where
$$\langle x^k, y^k \rangle = \sum_{\alpha\in\Gamma^k} x_\alpha^k y_\alpha^k \tag{4.4.10}$$

Let x^k be an element of U^k and $\varepsilon \in R$. Then we have

$$F(x^k+\varepsilon v^k) = \langle A^k(x^k+\varepsilon v^k) - 2r^k, x^k+\varepsilon v^k\rangle$$

$$= \langle A^k x^k - 2r^k, x^k\rangle + \varepsilon(\langle A^k x^k, v^k\rangle + \langle A^k v^k, x^k\rangle - 2\langle r^k, v^k\rangle) + \varepsilon^2\langle A^k v^k, v^k\rangle$$

Since A^k is symmetric i.e. $\langle A^k x^k, v^k\rangle = \langle A^k v^k, x^k\rangle$, we have

$$F(x^k+\varepsilon v^k) = F(x^k) + 2\varepsilon(\langle A^k x^k, v^k\rangle - \langle r^k, v^k\rangle) + \varepsilon^2\langle A^k v^k, v^k\rangle$$

It follows that

$$\frac{dF(x^k+\varepsilon v^k)}{d\varepsilon} = 2(\langle A^k x^k, v^k\rangle - \langle r^k, v^k\rangle) + 2\varepsilon\langle A^k v^k, v^k\rangle$$

and

$$\frac{d^2F(x^k+\varepsilon v^k)}{d\varepsilon^2} = 2\langle A^k v^k, v^k\rangle$$

A^k is positive definite implies

$$\left.\frac{d^2F(x^k+\varepsilon v^k)}{d\varepsilon^2}\right|_{\varepsilon=0} > 0 \qquad \text{if} \quad v^k \neq 0$$

Thus x^k minimizes F iff the first variation $\left.\frac{dF(x^k+\varepsilon v^k)}{d\varepsilon}\right|_{\varepsilon=0}$

vanishes for all v^k in U^k i.e.

$$\langle A^k x^k, v^k\rangle = \langle r^k, v^k\rangle \qquad \text{for all} \quad v^k \in U^k$$

Since the functions ϕ_α^k , $\alpha \in \Gamma^{k-1}$ form a basis for U^k, this holds for all v^k in U^k iff

$$<A^k x^k, \phi_\alpha^k> = <r^k, \phi_\alpha^k> \qquad \text{for all } \alpha \in \Gamma^{k-1}$$

$x^k \in U^k$ implies it can be written as

$$x^k = \sum_{\alpha \in \Gamma^{k-1}} x_\alpha^{k-1} \phi_\alpha^k = Q^k x^{k-1}$$

It follows that

$$<A^k Q^k x^{k-1}, \phi_\alpha^k> = <r^k, \phi_\alpha^k>$$

From (4.4.10) we have

$$\sum_{\beta \in \Gamma^k} (A^k Q^k x^{k-1})_\beta \phi_\alpha^k(\beta) = \sum_{\beta \in \Gamma^k} r_\beta^k \phi_\alpha^k(\beta)$$

From (4.4.3) we get

$$(P^k A^k Q^k x^{k-1})_\alpha = (P^k r^k)_\alpha \qquad \text{for all } \alpha \in \Gamma^{k-1}$$

It follows that

$$P^k A^k Q^k x^{k-1} = P^k r^k = r^{k-1}$$

From (4.4.4) we get

$$A^{k-1} x^{k-1} = r^{k-1}$$

i.e. A^{k-1} is the Ritz-Galerkin best approximation to A^k in the subspace U^k.

However, in general if A^k is not symmetric and positive definite, then the operator A^{k-1} can only be described as the Galerkin approximation to A^k.

4.5 THE ALGORITHM FAPIN

P.O. Frederickson [F4] introduced a new algorithm FAPIN to solve a large sparse linear systems of a certain class in $0(n)$ operations. In particular, it solves all finite element approximations, over a sufficiently regular mesh.

FAPIN is an iterative algorithm. At the beginning of the n^{th} pass one has an approximation x_n to the solution of $Ax = y$. An inner loop of FAPIN requires a 1-local ε-approximate inverse $C^k : Y^k \to X^k$ to A^k. If $Ax = y$ has a solution, Theorem 4.2.1 tells us that the initial vector x_0 can be random.

The iteration begins by computing the residual vector $r^{\ell} \leftarrow y - Ax_n$, continues by evaluating the residual vector r^k defined by (4.4.3) from r^{ℓ} to $r^{\ell 0}$, the residual vector at the bottom level $\ell 0$. Next, the approximate solution $z^{\ell 0} = C^{\ell 0} r^{\ell 0}$ is computed in the space $z^{\ell 0}$ and then one works back up from $k = \ell 0$ to $k = \ell -1$, first interpolating and then refining this approximation:

$$z^k \leftarrow Q^k z^{k-1}$$

$$z^k \leftarrow z^k + C^k(r^k - A^k z^k) \qquad (4.5.1)$$

At the top level, $k = \ell$, these assignments are replaced by

$$x_n^\ell \leftarrow x_n^\ell + Q^\ell z^{\ell-1}$$

$$x_{n+1}^\ell \leftarrow x_n^\ell + C^\ell(y - Ax_n^\ell) \qquad (4.5.2)$$

A detailed coding of the algorithm in Fortran to solve the linear system $Ax = y$ in a triangular domain is given in Appendix A.

The actual programs compute the norm of r^ℓ while computing r^ℓ and this is used to allow an early exit when tolerance ε has been achieved.

In general, if the operator A is not constant, then the lower approximations A^k must be computed first according to the equation (4.4.5). The corresponding approximate inverses C^k must also be evaluated. Techniques for construction of these approximate inverses will be discussed next.

4.6 CONSTRUCTION OF APPROXIMATE INVERSES

Benson [B3] has introduced several techniques to construct an approximate inverse for certain band matrices. In this section, we put the Truncation Technique (TRq) and Least-squares

Technique (LSq) [B3] into a slightly modified form and apply it to an 1-local linear operator A, to construct a 1-local operator C, an ε-approximate inverse to A. The TRq method is generalised by multiplying a weight W to the operator CA; we refer to this method as the Weighted Truncation Technique (WTq). However, approximate inverses obtain by these methods are not optimal. We introduce another new technique call Interpolation Technique (INq) to construct an optimal approximate inverse of A. This optimal inverse speeds up the convergence of the algorithm remarkably.

Denote by TRq(CA) the truncated q-local operator, where C and A are all q-local linear operators.

The TRq approximate inverse of A can be constructed by solving the system of linear equations

$$TRq(CA)_{\alpha,\beta} = \delta_{(0,0),\beta} \quad , \; |\beta|\le q \tag{4.6.1}$$

where δ denotes the Kronecker delta.

If A is the operator defined in (4.3.3) and the 1-local approximate inverse to A has the form (4.3.4), then it follows from (4.3.5) that the TRq approximate inverse C can be obtained by solving the following system of equations:

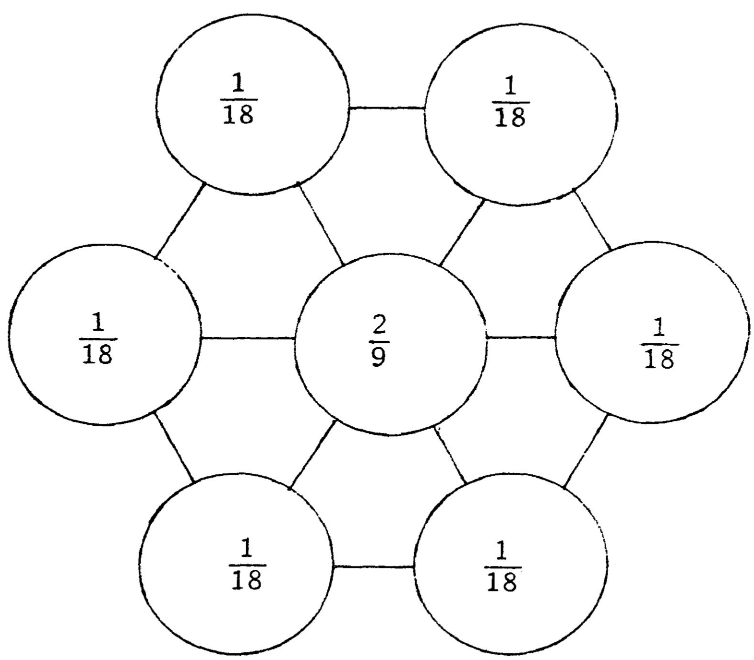
$$\begin{cases} a_1c_0 + (a_0+2a_1)c_1 = 0 \\ a_0c_0 + 6a_1c_1 = 1 \end{cases}$$

If $a_0^2 + 2a_0a_1 - 6a_1^2 \neq 0$, the above system of linear equations has an unique solution, i.e.

$$\begin{cases} c_0 = \dfrac{a_0+2a_1}{a_0^2+2a_0a_1-6a_1^2} \\[2ex] c_1 = \dfrac{a_1}{6a_1^2-2a_0a_1-a_0^2} \end{cases}$$

In particular, if A is the discrete Laplacian operator given in (4.4.6), then C has a representation of the form:

C:



Results with the TRq method applied to the discrete Laplacian operator A on a triangular domain at each level ℓ are tabulated below and graphically in Fig (4.6.4),

ℓ	n	ρ
2	15	0.3333
3	45	0.3591
4	153	0.4115
5	561	0.4612
6	2145	0.4757
7	8385	0.4751

where $n = (1+2^{\ell-1})(1+2^{\ell})$ is the total number of equations.

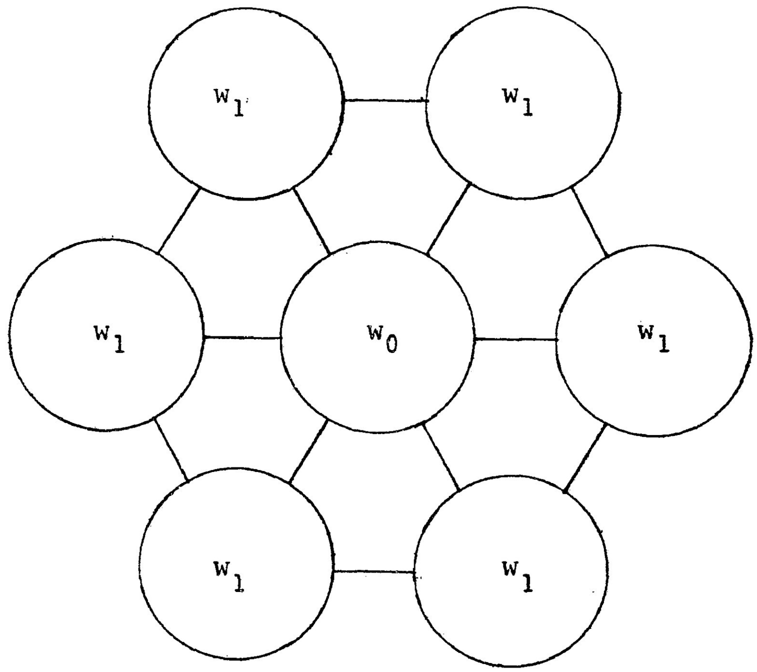
The TRq method can be generalized by multiplying a weight W to the operator CA, where W is a constant coefficient r-local operator.

The WTq approximate inverse C can be constructed by solving the system of linear equations

$$\mathrm{TRq}(CAW) = \mathrm{TRq}(W) \qquad (4.6.2)$$

If A and C are of the form (4.3.3) and (4.3.4) respectively, and W is a 1-local operator with a representation of the form

W:



then it follows from (4.3.5) and (4.3.2) that the linear system (4.6.2) becomes

$$\begin{cases} (a_0w_0+6a_1w_1)c_0+6[a_1w_0+(a_0+2a_1)w_1]c_1 = w_0 \\ [a_1w_0+(a_0+2a_1)w_1]c_0 + [(a_0+2a_1)w_0+(2a_0+15a_1)w_1]c_1 = w_1 \end{cases} \quad (4.6.3)$$

The linear system (4.6.3) always has an unique solution if

$$6[a_1w_0+(a_0+2a_1)w_1]^2 \neq (a_0w_0+6a_1w_1)[(a_0+2a_1)w_0+(2a_0+15a_1)w_1]$$

In particular, if W is chosen as A, then the linear system (4.6.3) becomes

$$\begin{cases} (a_0^2+6a_1^2)c_0 + 12a_1(a_0+a_1)c_1 = a_0 \\ 2a_1(a_0+a_1)c_0 + (a_0^2+4a_0a_1+15a_1^2)c_1 = a_1 \end{cases} \qquad (4.6.4)$$

We observe that the system (4.6.4) always has a solution. If A is the discrete Laplacian operator, we have

$$\begin{cases} c_0 = 17/89 = 0.1910112 \\ c_1 = 3/89 = 0.0337079 \end{cases} \qquad (4.6.5)$$

Results with the WTq method applied to the discrete Laplacian operator A on a triangular domain at each level ℓ are tabulated below and graphically in Fig (4.6.4),

ℓ	n	ρ
2	15	0.1011
3	45	0.1461
4	153	0.1510
5	561	0.1698
6	2145	0.1746
7	8385	0.1748

Denote by $\|\cdot\|_{\alpha,2} = (\sum\limits_{|\beta|\leq q} a^2_{\alpha,\beta})^{\frac{1}{2}}$ the discrete ℓ^2 norm of the q-local operator A at the point $\alpha \in \mathring{\Gamma}_h$, by $g_\alpha = \|\cdot\|^2_{\alpha,2}$. If C and A are the linear operators defined in (4.3.4) and (4.3.3) respectively, then

$$g_\alpha = \| CA-I \|^2_{\alpha,2} = (a_0c_0+6a_1c_1-1)^2 + 6[a_1c_0+(a_0+2a_1)c_1]^2 + 30(a_1c_1)^2 \tag{4.6.6}$$

We observe that g_α is a function of the parameters c_0 and c_1, the methods of calculus enable us to find the values of c_0 and c_1 that minimize g. The approximate inverse C obtained by this method is called the LSq approximate inverse and we refer to this technique as the LSq method.

From (4.6.6) we have

$$\frac{\partial g_\alpha}{\partial c_0} = 2(a_0c_0+6a_1c_1-1)c_0 + 12[a_1c_0+(a_0+2a_1)c_1]a_1$$

$$\frac{\partial g_\alpha}{\partial c_1} = 2(a_0c_0+6a_1c_1-1)(6a_1) + 12[a_1c_0+(a_0+2a_1)c_1](a_0+2a_1)+60a_1^2c_1$$

To minimize g_α, we require $\frac{\partial g_\alpha}{\partial c_0} = 0$ and $\frac{\partial g_\alpha}{\partial c_1} = 0$, i.e.

$$\begin{cases} (a_0^2+6a_1^2)c_0 + 12a_1(a_0+a_1)c_1 = a_0 \\ \\ 2a_1(a_0+a_1)c_0 + (a_0^2+4a_0a_1+15a_1^2)c_1 = a_1 \end{cases}$$

We observe that the above system of linear equations turns out to be the same as the WTq method applied to same operator CA with a weight $W = A$.

In general, if the six coefficients $a_{\alpha,\beta}$, $|\beta|=1$ are not equal, then the approximate inverse C at each point $\alpha \in \mathring{\Gamma}_h$ has 7 parameters to be determined. It follows from (4.3.2) that

$$g_\alpha = \|CA-I\|^2_{\alpha,2} = \Big(\sum_{|\beta|\le 1} c_{\alpha,\beta}a_{\alpha+\beta,-\beta}-1\Big)^2 + \sum_{1\le|\gamma|\le 2}\Big(\sum_{\substack{\beta \\ |\beta|\le 1 \\ |\gamma-\beta|\le 1}} c_{\alpha,\beta}a_{\alpha+\beta,\gamma-\beta}\Big)^2$$

To minimize g_α, we require $\dfrac{\partial g_\alpha}{\partial c_{\alpha,\sigma}} = 0$ for $\sigma \in \Gamma_h$, $|\sigma-\alpha|\le 1$. Now

$$\frac{\partial g_\alpha}{\partial c_{\alpha,\sigma}} = 2\Big(\sum_{|\beta|\le 1} c_{\alpha,\beta}a_{\alpha+\beta,-\beta}-1\Big)a_{\alpha+\sigma,-\sigma} + \sum_{\substack{\gamma \\ 1\le|\gamma|\le 2, \\ |\gamma-\sigma|\le 1}} 2\Big(\sum_{\substack{\beta \\ |\beta|\le 1 \\ |\gamma-\beta|\le 1}} c_{\alpha,\beta}a_{\alpha+\beta,\gamma-\beta}\Big)a_{\alpha+\sigma,\gamma-\sigma} = 0$$

which gives

$$a_{\alpha+\sigma,-\sigma}\Big(\sum_{|\beta|\le 1} c_{\alpha,\beta}a_{\alpha+\beta,-\beta}-1\Big) + \sum_{\substack{\gamma \\ 1\le|\gamma|\le 2 \\ |\gamma-\sigma|\le 1}} a_{\alpha+\sigma,\gamma-\sigma} \sum_{\substack{\beta \\ |\beta|\le 1 \\ |\gamma-\beta|\le 1}} c_{\alpha,\beta}a_{\alpha+\beta,\gamma-\beta} = 0$$

for $\sigma \in \Gamma_h$, $|\sigma-\alpha|\le 1$.

Thus, the 1-local LSq-approximate inverse of A at the point $\alpha \in \mathring{\Gamma}_h$ can be obtained by solving the above linear system of 7 equations. This linear system of equations can also be written as

$$\sum_{|\beta|\le 1} c_{\alpha,\beta} \sum_{\substack{\gamma \\ |\gamma|\le 2,\ |\gamma-\sigma|\le 1 \\ |\gamma-\beta|\le 1}} a_{\alpha+\sigma,\gamma-\sigma}a_{\alpha+\beta,\gamma-\beta} = a_{\alpha+\gamma,-\sigma} \qquad (4.6.7)$$

for $\sigma \in \Gamma_h$, $|\sigma-\alpha|\le 1$

The sum extends over only those γ for which $\gamma-\sigma$, $\gamma-\beta$ $\in \Gamma_h$.

If A is a constant coefficient 1-local operator, we are also interested to construct an approximate inverse of A by the application of LSq method to the weighted operator ACA, and try to minimize the expression

$$g_\alpha = \|ACA - A\|^2_{\alpha,2}$$

It follows from (4.3.5) and (4.3.2) that

$$g_\alpha = [(a_0^2+6a_1^2)c_0+12a_1(a_0+a_1)c_1+a_0]^2 + 6[2a_1(a_0+a_1)c_0+(a_0^2+4a_0a_1+15a_1^2)c_1-a_1]^2$$

$$+ 6[a_1^2c_0+2a_1(a_0+3a_1)c_1]^2 + 6[2a_1^2c_0+2a_1(2a_0+3a_1)c_1]^2 + 114a_1^4c_1^2$$

Let $\dfrac{\partial g_\alpha}{\partial c_0} = 0$ and $\dfrac{\partial g_\alpha}{\partial c_1} = 0$ we have

$$\begin{cases} (a_0^4+36a_0^2a_1^2+48a_0a_1^3+90a_1^4)c_0 + 24a_1(a_0^3+3a_0^2a_1+15a_0a_1^2+15a_1^3)c_1 \\ \qquad = a_0^3+18a_0a_1^2+12a_1^3 \\ 4a_1(a_0^3+3a_0^2a_1+15a_0a_1^2+15a_1^3)c_0 + (a_0^4+8a_0^3a_1+90a_0^2a_1^2+240a_0a_1^3+340a_1^4)c_1 \\ \qquad = 3a_0^2a_1+6a_0a_1^2+15a_1^3 \end{cases}$$

In particular, if A is the discrete Laplacian operator, then we have

$$\begin{cases} c_0 = 103/597 = 0.1725293 \\ c_1 = 1117/48556 = 0.0230044 \end{cases} \qquad (4.6.8)$$

Application of the above approximate inverse to the algorithm FAPIN on a triangular domain, the numerical results are tabulated below and graphically in Fig (4.6.4).

ℓ	n	ρ
2	15	0.1258
3	45	0.1539
4	153	0.1691
5	561	0.1714
6	2145	0.1672
7	8385	0.1637

The approximate inverse C determined by the TRq or LSq method is usually not optimal,however, it can be improved by the INq method. This method is feasible only in the constant coefficient case. For simplicity, we shall introduce this technique with an example for the construction of an optimal ε-approximate inverse to the discrete Laplacian operator A.

Let $\tilde{c}_0$ and $\tilde{c}_1$ be two approximate parameters of the operator C obtain by TRq or LSq method. Then the optimal values of c_0 and c_1 can be obtained by the following steps:

Step I. $\tilde{c}_0$ is held fixed. Perturbing c_1 about the point $\tilde{c}_1$, we obtain a set of experimental data (ρ,c_1). The point where ρ has a minimum can be obtained by plotting the graph of ρ against c_1.

Step II. Perturbing c_0 about $\tilde{c}_0$, for each fixed values of c_0, carry out the same procedures as in Step I to obtain a set of points $(c_0, c_1^{opt}, \rho_{min})$.

Step III. c_1 is held fixed instead of c_0, repeating the whole procedures as in Step I and II, we obtain another set of points $(c_0^{opt}, c_1, \rho_{min})$.

Step IV. Plotting the graph of c_1 against c_0 for the data (c_0, c_1^{opt}) and (c_0^{opt}, c_1) collected in Step II and III, we find that the curves intersect at a point (c_0^{opt}, c_0^{opt}), this is the optimal solution of the operator C.

To illustrate the method, three graphs of c_1 against c_0 for the data collected in Step II and III of the INq method at level ℓ = 2,3 and 4 are plotted in Fig (4.6.1), Fig (4.6.2) and Fig (4.6.3) respectively. In order to have a clear picture of the behaviour of ρ near the optimal solution (c_0^{opt}, c_1^{opt}), three contour graphs of ρ at different height are also plotted in these graphs.

The INq ε-approximate inverses C at level ℓ = 2,3 and 4 are shown in Table 4.6.1. Application of these INq ε-approximate inverses C to the algorithm FAPIN, the spectral radius of the operator I-AC at each level ℓ are shown in Table 4.6.2 and graphically in Fig 4.6.4.

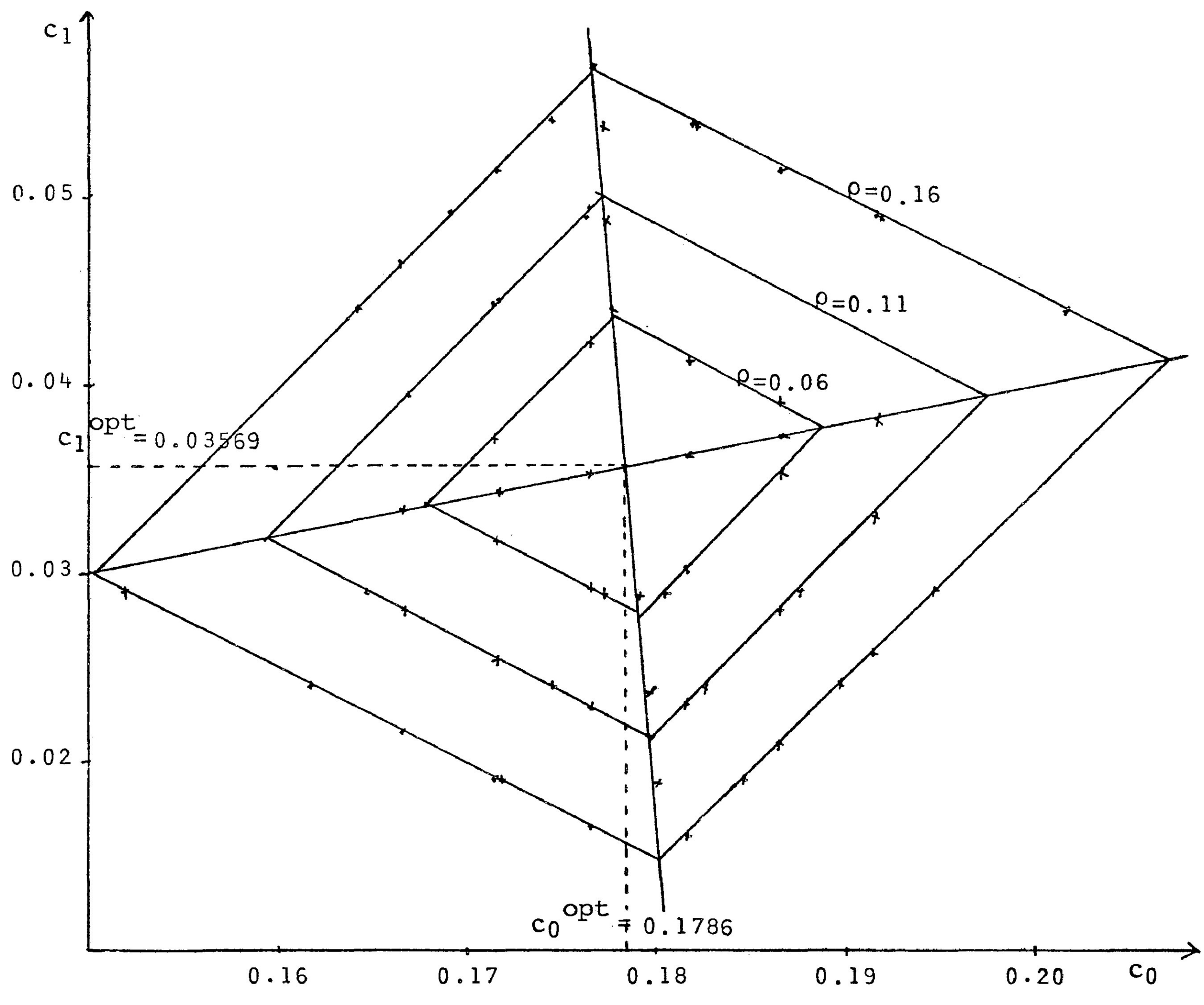
Table 4.6.1

ℓ	c_0^{opt}	c_1^{opt}
2	0.1786	0.03569
3	0.1803	0.02921
4	0.1825	0.02791

Table 4.6.2

ℓ	n	ρ		
		opt. level ℓ = 2	opt. level ℓ = 3	opt. level ℓ = 4
2	15	0.0003	0.0578	0.0821
3	45	0.2224	0.0822	0.0950
4	153	0.3072	0.1370	0.1037
5	561	0.3318	0.1510	0.1001
6	2145	0.3368	0.1528	0.1190
7	8385	0.3326	0.1532	0.1310

Fig. 4.6.1
(INq method, ℓ=2)



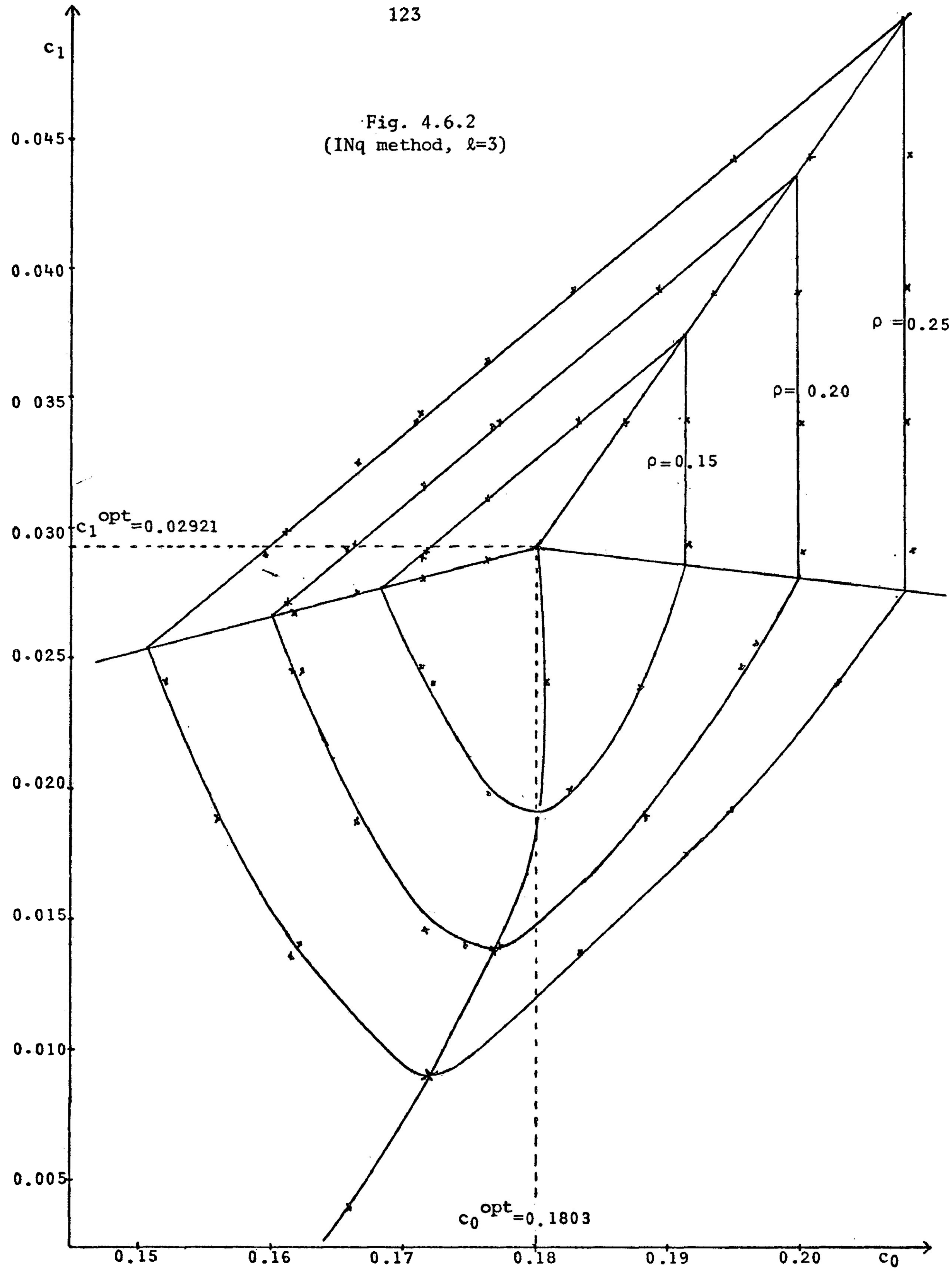
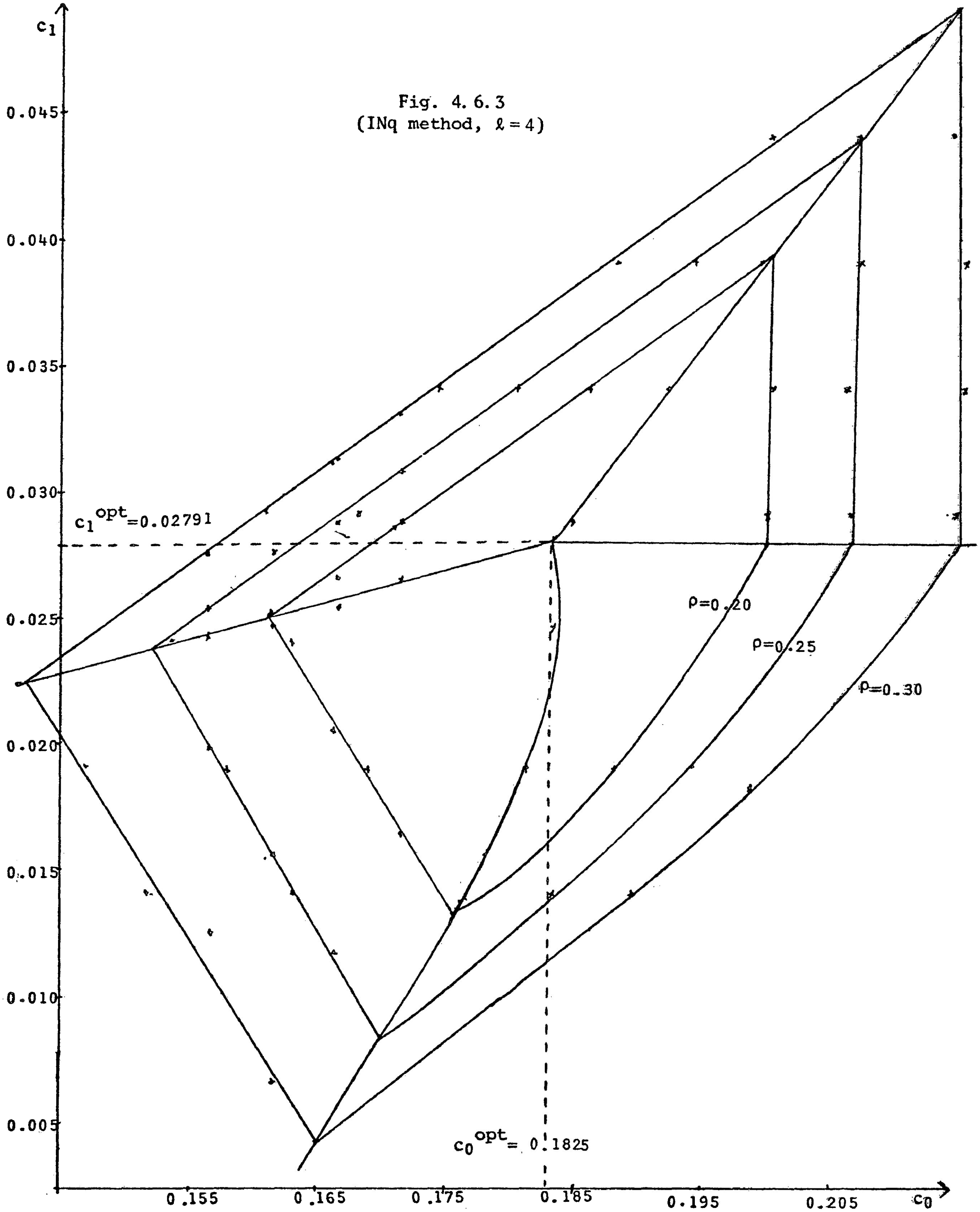


Fig. 4.6.2
(INq method, ℓ=3)
c1
0.045
0.040
0 035
0.030
0.025
0.020
0.015
0.010
0.005
c1opt=0.02921
ρ = 0.25
ρ= 0.20
ρ=0.15
c0opt=0.1803
0.15
0.16
0.17
0.18
0.19
0.20
c0


c1
0.045
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
Fig. 4.6.3
(INq method, ℓ = 4)
c1opt=0.02791
ρ=0.20
ρ=0.25
ρ=0.30
c0opt= 0.1825
0.155
0.165
0.175
0.185
0.195
0.205
c0

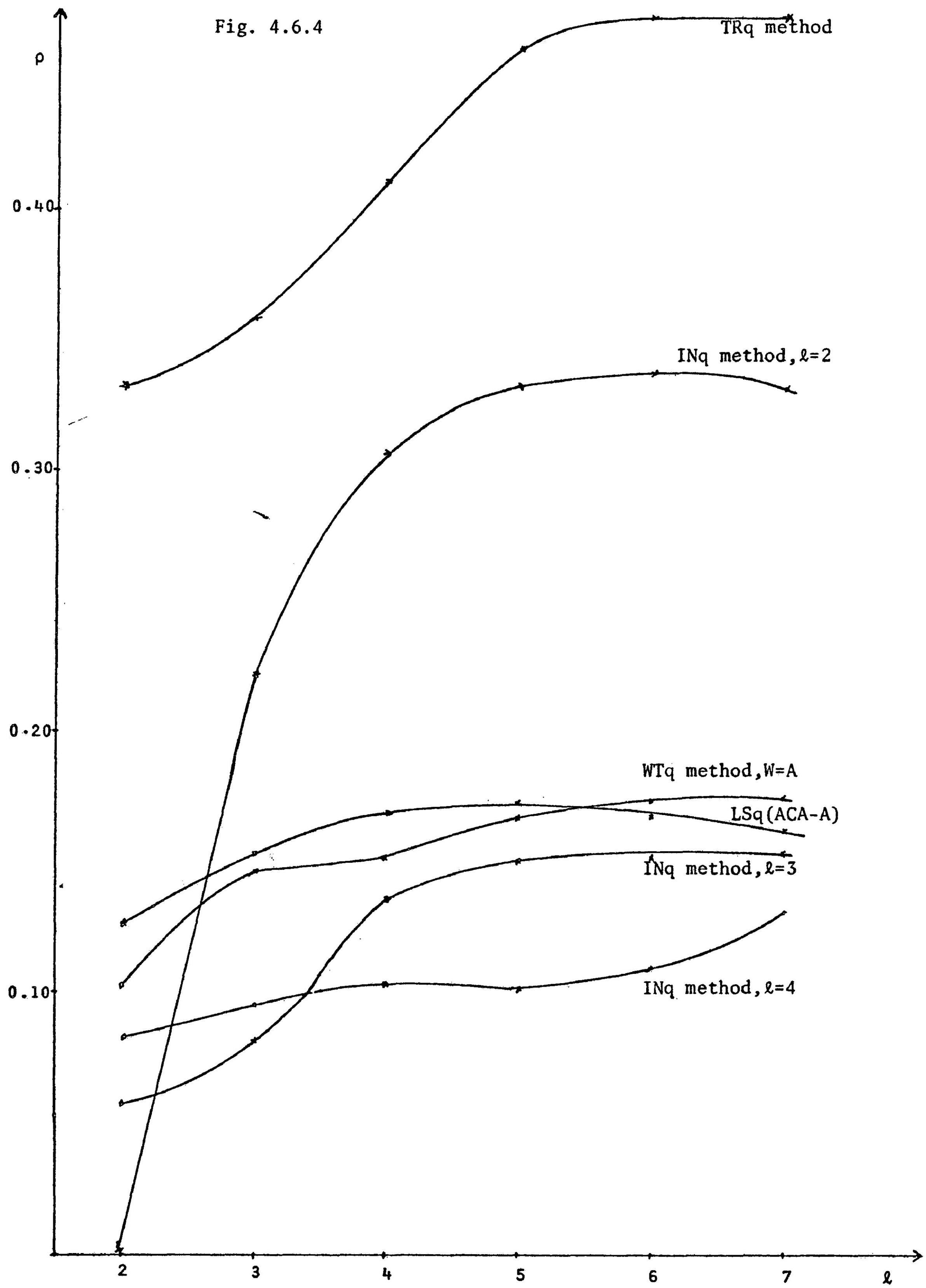


Fig. 4.6.4
ρ
0.40
0.30
0.20
0.10
TRq method
INq method,ℓ=2
WTq method,W=A
LSq(ACA-A)
INq method,ℓ=3
INq method,ℓ=4
2
3
4
5
6
7
ℓ

From the experimental results, we observe that the INq ε-approximate Inverse C varies from one level to another level, they are only optimal at the constructed level. In order to have a clear picture of the behaviour of the spectral radius as ℓ becomes large, a chart of the spectral radius ρ against ℓ for the various construction techniques are plotted in Fig 4.6.4.

As we can see from the graphs in Fig 4.6.4, the rate of convergence is independent of n for equations in the class considered. When ℓ becomes large, the spectral radius of I-CA, ρ tends to a certain value.

We observe that the spectral radius $\rho(I-AC)$ for C constructed by the LSq method or WTq method with weight $W = A$ are not too far away from its optimal value. We are interested to know what is the best choice of the weight W, to make the WTq approximate inverse becomes optimal?

If W is a 1-local operator, from (4.6.3), we have

$$\begin{cases} (a_0c_0+6a_1c_1-1)w_0 + 6[a_1c_0+(a_0+2a_1)c_1]w_1 = 0 \\ [a_1c_0+(a_0+2a_1)c_1]w_0 + [(a_0+2a_1)c_0+(2a_0+15a_1)c_1-1]w_1 = 0 \end{cases} \tag{4.6.12}$$

The linear system (4.6.12) has non-trivial solutions iff

$$6[a_1c_0+(a_0+2a_1)c_1]^2 = (a_0c_0+6a_1c_1-1)[(a_0+2a_1)c_0+(2a_0+15a_1)c_1-1]$$

It follows that (c_0,c_1) are related by

$$(6a_1^2-2a_0a_1-a_0^2)c_0^2 - (2a_0^2+9a_0a_1-12a_1^2)c_0c_1 + 6(a_0^2+2a_0a_1-11a_1^2)c_1^2 +$$
$$2(a_0+a_1)c_0 + (2a_0+21a_1)c_1 - 1 = 0$$

In particular, if A is the discrete Laplacian operator $a_0 = 6$, $a_1 = -1$, we have

$$78c_1^2 - 6c_0c_1 - 18c_0^2 + 10c_0 - 9c_1 - 1 = 0 \qquad (4.6.13)$$

The locus of the above equation is a hyperbola with $c_0 \geq 0.3047298$ or $c_0 \leq 0.2281788$.

The LSq(ACA) ε-approximate inverse obtain in (4.6.8) and the INq ε-approximate inverses at level $\ell = 2,3$ and 4 cannot fix into the equation (4.6.13) exactly. For each c_1 we have constructed before, the corresponding $\tilde{c}_0$ obtain from (4.6.13) which is closest to those constructed value c_0 and the corresponding weight W are tabulated below:

Construction Technique	constructed		From (4.6.13)	Weight W	
	c_0	c_1	$\tilde{c}_0$	w_0	w_1
TRq	0.2222222	0.0555556	0.2222222	1	0
LSq(AC-I)	0.1910112	0.0337079	0.1910112	6	-1
LSq(ACA-A)	0.1725293	0.0230044	0.1725503	4.704	-1
INq, ℓ=2	0.1786	0.03569	0.1943	6.400	-1
INq, ℓ=3	0.1803	0.02921	0.1833	5.322	-1
INq, ℓ=4	0.1825	0.02791	0.1811	5.169	-1

4.7 EXPERIMENTAL RESULTS

We now discuss some numerical examples of boundary value problems, whose solutions have been approximated by the Ritz-Galerkin approximation discussed in Chapter 2.

Consider the problem

$$\begin{cases} Lu = -\Delta u(x_0,x_1,x_2) = \frac{4}{3}\sum_i \sin(1-2x_i) & \text{in } \Omega \\ u = 0 & \text{on } \Omega \end{cases} \qquad (4.7.1)$$

where Ω is an equilateral triangle of unit side length, and (x_0,x_1,x_2) is the Barycentric Coordinates of a point X in the triangle Ω.

The unique solution to (4.7.1) is

$$u(x_0,x_1,x_2) = \sin(x_0)\sin(x_1)\sin(x_2)$$

The solution of (4.7.1) was approximated by minimizing the quadratic functional

$$I(u) = \int_{\Omega} [\nabla u \cdot \nabla u - \frac{8}{3} u \sum_i \sin(1-2x_i)]\, d\mu_\Omega$$

over the piecewise linear subspace $S_0^{1,0}$ of $H_0^1(\Omega)$.

It follows from (2.5.4) and (2.5.6) that we are solving the 1-local linear system

$$Au^h = \frac{3h^2}{4\mu_\Omega(T)} \int_\Omega f\phi_\alpha d\mu_\Omega = \frac{h^2}{\mu_\Omega(T)} \int_\Omega \tilde{f}\phi_\alpha d\mu_\Omega$$

where A is the discrete Laplacian operator defined in (4.4.6) and $\tilde{f} = \sum_i \sin(1-2x_i)$

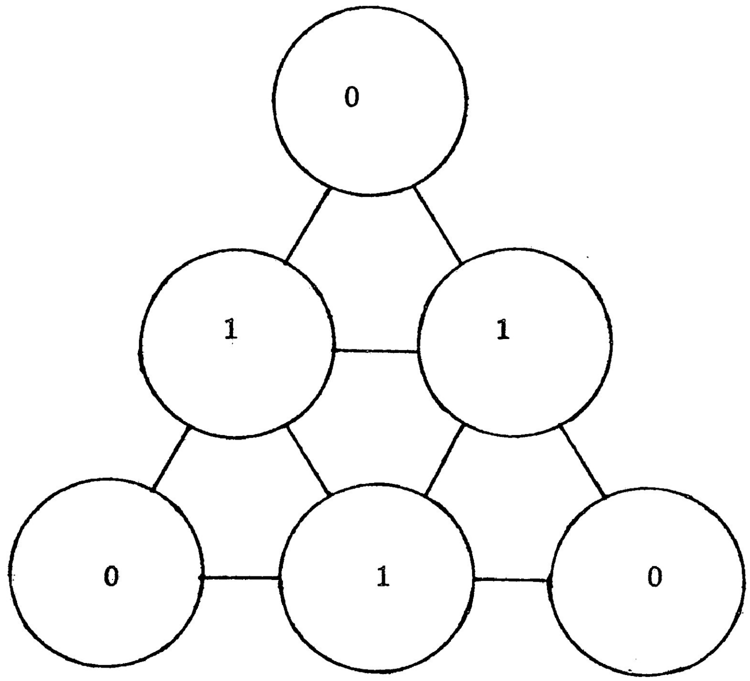
If the 1-point numerical quadrature is used, then we are solving the linear system

$$A\tilde{u}^h = 2h^2\tilde{f}(X_\alpha)$$

The numerical results are given in Table 4.7.1. The quantity s in this table is

$$s = \log\left(\frac{\|u-\tilde{u}^{h_1}\|_{L^2(\Omega)}}{\|u-\tilde{u}^{h_2}\|_{L^2(\Omega)}}\right) / \log\left(\frac{h_1}{h_2}\right)$$

The norm $\|u-\tilde{u}^h\|_{L^2(\Omega)}$ is approximated by applying some numerical quadrature to each of the triangular elements $T \in \tau^h$. In our numerical experiment, the third order Gregory type formula [L1,p74] are used to approximate the norm $\|u-\tilde{u}^h\|_{L^2(\Omega)}$.



We see from Table 4.7.1 that the accuracy seems to be $O(h^2)$ in the norm $\|\cdot\|_{L^2(\Omega)}$

Table 4.7.1 (1-point formula)

ℓ	h	$\|u-\tilde{u}^h\|_{L^2(\Omega)}$	s
2	0.25	5.4427×10^{-3}	
3	0.125	1.3725×10^{-3}	1.99
4	0.0625	3.4390×10^{-4}	2.00
5	0.03125	8.6298×10^{-5}	2.00

If the 7-point numerical quadrature is used, then $v_\alpha^h = h^2 \sum_{|\beta|\le 1} F_{\alpha,\beta}$ where $F_{\alpha,\beta}$ is the 7-point numerical quadrature apply to the function

$$f = \sum_i \sin(1-2x_i)$$

The numerical results for the 7-point numerical quadrature are given in Table 4.7.2

We see from the Table 4.7.2 that the accuracy seems to be $O(h^2)$ in the norm $\|\cdot\|_{L^2(\Omega)}$

Table 4.7.2 (7-point formula)

ℓ	h	$\|\|u-\tilde{u}^h\|\|_{L^2(\Omega)}$	s
2	0.25	5.6333×10^{-3}	
3	0.125	1.4339×10^{-3}	1.97
4	0.0625	3.6016×10^{-4}	1.99
5	0.03125	9.0370×10^{-5}	2.00

Our second example is the problem of inhomogenous boundary condition defined by

$$\begin{cases} Lu = -\Delta u(x_0,x_1,x_2) = \frac{8}{3}\sum_i (1-2x_i)e^{-x_i^2} & \text{in } \Omega \\ u(x_0,x_1,x_2) = \sum_i e^{-x_i^2} & \text{on } \partial\Omega \end{cases} \tag{4.7.2}$$

where Ω is an equilateral triangle of unit side length. The unique solution to (4.7.2) is

$$u(x_0,x_1,x_2) = \sum_i e^{-x_i^2}$$

The Ritz-Galerkin approximation to the problem (4.7.2) in the finite dimensional affine space $S_g^{1,0}$ yields the following system of linear equations:

$$Au^h = \frac{h^2}{\mu_\Omega(T)}\int_\Omega f\phi_\alpha d\mu_\Omega$$

where $f = 2\sum_i (1-2x_i)e^{-x_i^2}$ and A is the discrete Laplacian operator.

If the 1-point or 7-point numerical quadrature is used, we are solving the following 1-local linear system

$$A\tilde{u}^h = F^h$$

This linear system can be solved by the algorithm FAPIN as easy as the homogeneous boundary condition case by simply pre-set the values of $\tilde{u}^h$ on the boundary of Ω_h by $\sum_i e^{-x_i^2}$ instead of zeros.

The results of the 1-point and 7-point numerical quadratures are given in Table 4.7.3 and Table 4.7.4 respectively. It seems from the results in these tables that the accuracy of the Ritz-Galerkin solution to the problem (4.7.2) are probably $O(h^2)$.

Table 4.7.3 (1-point formula)

ℓ	h	$\|\|u-\tilde{u}^h\|\|_{L^2(\Omega)}$	s
2	0.25	1.9278×10^{-2}	
3	0.125	4.7939×10^{-3}	2.01
4	0.0625	1.1945×10^{-3}	2.00
5	0.03125	2.8678×10^{-4}	2.06

Table 4.7.4 (7-point formula)

ℓ	h	$\|\|u-\tilde{u}^h\|\|_{L^2(\Omega)}$	s
2	0.25	2.0306×10^{-2}	
3	0.125	5.1281×10^{-3}	1.99
4	0.0625	1.2849×10^{-3}	2.00
5	0.03125	3.1192×10^{-4}	2.04

As we can see from the first two examples, although the 7-point formula is more accurate than the 1-point formula, when they are applied to the Ritz-Galerkin approximation, for certain types of function u, the error in the 1-point formula may cancel off part of the error induced by the Ritz-Galerkin approximation and give a better approximation to the true solution u than using the 7-point formula would give.

Our last example is to apply the algorithm FAPIN to solve the problem.

$$\begin{cases} Lu = -\Delta u + \lambda u = f & \text{in } \Omega \\ u = \sin(x_0)\sin(x_1)\sin(x_0 - x_1) & \text{on } \partial\Omega \end{cases} \qquad (4.7.3)$$

with f chosen to be Lu and

$$u = \sin(x_1 - x_2)\sin(x_1)\sin(x_2)$$

where Ω is an equilateral triangle of unit side length, and λ is equal to one of the eigenvalues of the operator $\Delta u = \lambda u$.

If $u_\lambda = \sin(2\pi x_0) + \sin(2\pi x_1) + \sin(2\pi x_2)$, then it is easy to check that $u_\lambda = 0$ on $\partial\Omega$.

For this function u, we have

$$D_i u_\lambda = -2\pi \mathrm{Cos}(2\pi x_{i+1}) + 2\pi \mathrm{Cos}(2\pi x_{i-1})$$

$$D_{i,i} u_\lambda = -4\pi^2 \sin(2\pi x_{i+1}) - 4\pi^2 \sin(2\pi x_{i-1})$$

It follows that

$$\Delta u_\lambda = \frac{2}{3}\sum_i D_{i,i} u = -\frac{16\pi^2}{3}\sum_i \sin(2\pi x_i) = -\frac{16\pi^2}{3} u$$

Thus $\lambda = -\frac{16\pi^2}{3}$ is the eigenvalue corresponding to the eigenfunction $u_\lambda = \sum_i \sin(2\pi x_i)$ of the Laplacian operator Δ.

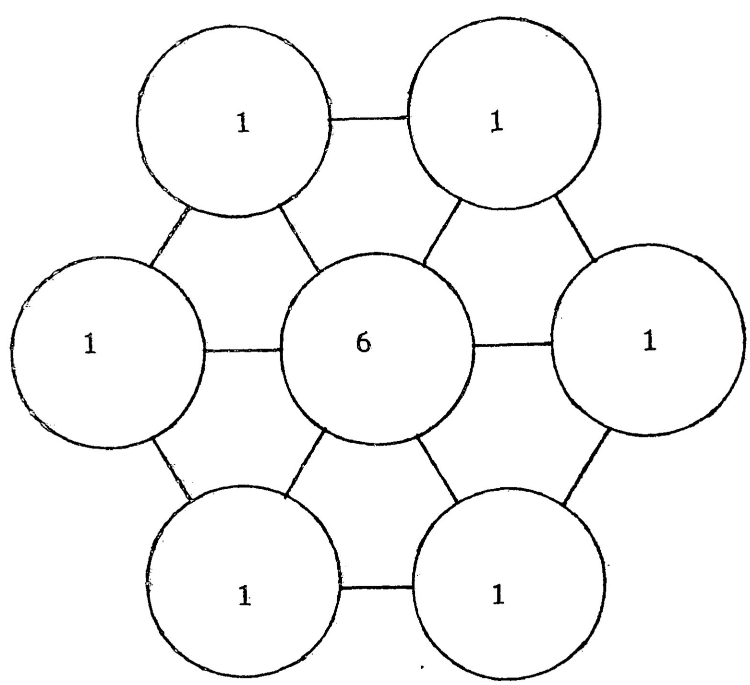
In fact, $\lambda_n = -\frac{16n^2\pi^2}{3}$ are the eigenvalues corresponding to the eigenfunctions $u_\lambda = \sum_i \sin(2n\pi x_i)$ for all $n \in N$

When $\lambda = -\frac{16\pi^2}{3}$, the operator $L = -\Delta + \lambda I$ is singular. It follows from (2.5.5) that the Ritz-Galerkin solutions to (4.7.3) is the solution of the following linear system

$$L^h u^h = (A^h + \lambda B^h) u^h = \frac{3h^2}{4} \int_\Omega f \phi_\alpha d\mu_\Omega \tag{4.7.4}$$

where A^h is the discrete Laplacian operator and $B^h = \frac{h^2}{8} \tilde{B}^h$, $\tilde{B}^h$ can be represented as

$\tilde{B}^h$:



If the 1-point or 7-point numerical quadrature is used, we are solving the linear system

$$L^h\tilde{u}^h = (A^h+\lambda B^h)\tilde{u}^h = F^h \qquad (4.7.5)$$

In this case, $\lambda = -\frac{16\pi^2}{3}$ is approximately equal to the discrete eigenvalue λ^h of L^h. Thus the linear operator L^h is nearly singular. The linear system (4.7.5) becomes difficult to solve by some algorithm. However, if (4.7.5) has a solution. Theorem 4.2.1 tells us that a solution to (4.7.5) is constructed by (4.2.5).

Since the problem (4.7.3) has a solution

$$u = \sin(x_0)\sin(x_1)\sin(x_0-x_1)$$

thus the linear system (4.7.5) still can be solved by the algorithm FAPIN, although L^h is almost singular.

It follows from (4.4.8) that the Ritz-Galerkin best approximation to the operator L^k at the k^{th} level can be written as

$$\begin{cases} L^\ell = A^h + \lambda B^h & \\ L^{k-1} = A^k + 4\lambda B^k & \text{for} \quad 2\le k\le \ell \end{cases}$$

or they can be expressed in terms of A^h and $\tilde{B}^h$ as

$$L^k = A^h + 4^{\ell-k}\left(\frac{\lambda h^2}{8}\right)\tilde{B}^h \qquad \text{for} \quad 2\le k\le \ell$$

The approximate inverse for L^k at level k can be constructed by the WTq method for a proper choice of weight W.

If u is a solution to the equation (4.7.3), since L is singular, it implies $u + \kappa u_\lambda$ is also a solution to $Lu = f$, where κ is a constant and u_λ is the eigenfunction of Δ corresponding to the eigenvalue λ. Because of the symmetry of the algorithm we are using, the solution $\tilde{u}^h$ is, like $\tilde{F}$, antisymmetric with respect to the line $x_1 - x_2 = 0$. Thus $\kappa = 0$, and we are able to compare $\tilde{u}^h$ with u.

Numerical results with the 1-point and 7-point formulas apply to (4.7.4) are given in Table 4.7.5 and Table 4.7.6 respectively. It seems from these tables that the accuracy of the Ritz-Galerkin solutions to the problem (4.7.3) for the 1-point and 7-point numerical quadratures are both $O(h^2)$.

Table 4.7.5 (1-point formula)

ℓ	h	$\|\|u-\tilde{u}^h\|\|_{L^2(\Omega)}$	s
2	0.25	1.5042×10^{-2}	
3	0.125	3.3667×10^{-3}	2.16
4	0.0625	8.9534×10^{-4}	1.91
5	0.03125	2.2858×10^{-4}	1.97

Table 4.7.6 (7-point formula)

ℓ	h	$\|u-\tilde{u}^h\|_{L^2(\Omega)}$	s
2	0.25	1.8307×10^{-2}	
3	0.125	2.9901×10^{-3}	2.61
4	0.0625	7.7986×10^{-4}	1.94
5	0.03125	1.9813×10^{-4}	1.98

An even more striking demonstration is provided by taking $\lambda = \lambda^h$, in this case the linear operator L^h is almost singular, and yet the linear system still can be solved by the algorithm FAPIN.

Numerical results for $\lambda = \lambda^h = -52.810$ at level $\ell = 5$ are given in Table 4.7.7. The norm $\|F^h - L^h\tilde{u}_k^h\|_2$ in this table is the root-mean-square of the residual $F^h - L^h\tilde{u}_k^h$.

Table 4.7.7 ($L^h u^h = (A^h + \lambda^h B^h)u^h = F^h$, $u_0 = 0$)

Iteration	$\|F^h - L^h u_k^h\|_2$, $\lambda^h = -52.810$	
	1-point formula	7-point formula
0	3.0428×10^{-2}	3.0430×10^{-2}
1	4.6462×10^{-3}	4.6464×10^{-3}
2	3.5905×10^{-4}	3.5907×10^{-4}
3	2.8968×10^{-5}	2.8963×10^{-5}
4	3.5354×10^{-6}	3.5372×10^{-6}
5	3.8152×10^{-7}	3.7858×10^{-7}
6	1.2493×10^{-7}	1.2448×10^{-7}
7	6.4122×10^{-8}	6.4493×10^{-8}

The rate of convergence for the 1-point and 7-point formula with $\lambda^h = -52.810$ are showed in Table 4.7.8 and Table 4.7.9 respectively.

Table 4.7.8 (1-point formula, $\lambda^h = -52.810$)

ℓ	h	$\|u-\tilde{u}^h\|_{L^2(\Omega)}$	s
2	0.25	1.5626×10^{-2}	
3	0.125	3.3356×10^{-3}	2.23
4	0.0625	8.5449×10^{-4}	1.96
5	0.03125	1.8673×10^{-4}	2.19

Table 4.7.9 (7-point formula, $\lambda^h = -52.810$)

ℓ	h	$\|u-\tilde{u}^h\|_{L^2(\Omega)}$	s
2	0.25	1.9033×10^{-2}	
3	0.125	2.9610×10^{-3}	2.68
4	0.0625	7.4242×10^{-4}	2.00
5	0.03125	1.6120×10^{-4}	2.20

We observe that as ℓ becomes large, the vector F^h in the linear system $L^h u^h = F^h$ tends to zero and u^h tends to the exact solution u. But in terms of actual computing, because of the round off error, the Ritz-Galerkin solution to the problem

$Lu = f$ can only give a good approximation in single arithmetic if the level ℓ is less than 6. However, a better approximation can be obtained by refining the mesh and using the double precision arithmetic.

APPENDIX A

FORTRAN PROGRAMS OF FAPIN FOR SOLVING A 1-LOCAL LINEAR SYSTEM IN A TRIANGULAR DOMAIN

In this appendix, we describe in detail the FORTRAN subroutine FAPIN for solving a 1-local linear system $Ax = y$ in a triangular domain Ω.

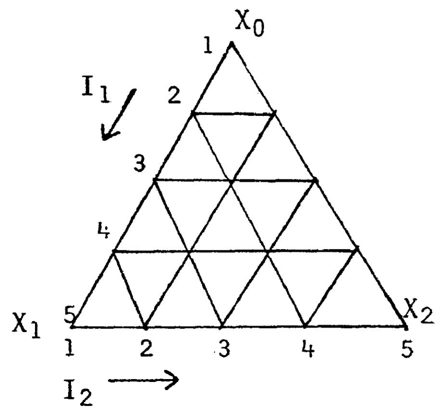


Fig. A1

As shown in Fig. A1, the integer lattice (i_1,i_2) of the triangular grids are numbered from top to bottom for i_1 and from left to right for i_2. The vectors x^k, y and r^k are all stored in each of the one dimensional array X, Y and R respectively. In particular, we store $x^k_{i_1,i_2}$ as X(N(K)+M(I1)+I2), $r^k_{i_1,i_2}$ as R(N(K)+M(I1)+I2), $y^{\ell}_{i_1,i_2}$ as Y(M(I1)+I2).

Starting with N(1) = 0, M(I1) represents the total number of points in row 1, row 2, $\cdots$ up to row (i_1-1). Similarly, with N(L) = 0, N(K) indicates the total number of points in Γ^{ℓ}, $\Gamma^{\ell-1}$, $\cdots$ up to Γ^{k-1}.

In each of the iteration, the residual vector $r^{\ell} \leftarrow y-Ax$ and $r^k \leftarrow r^k-A^kx^k$ are computed in the subroutine RESINV by setting the logical parameter RESIDU = ·TRUE·, the vectors $x^k \leftarrow x^k+B^k(r^k)$ are also evaluated in this subroutine by setting

RESIDU = .FALSE. The projection steps $r^{k-1} \leftarrow P^k(r^k)$ and interpolation steps $x^k \leftarrow Q^k(x^{k-1})$ are carried out in the subroutine FAPIN. Once the norm $\|r\|$ is less than the tolerance TOL or when the number of iterations reaches NIT-1, the computed results are passed to the calling program.

Fig. A1

```
C WHEN RESIDU = .TRUE.,TO COMPUTE THE RESIDUAL VECTOR RK=RK-AK(XK)
C      THE TWO CONSTANT COEFFICIENTS OF -AK ARE STORED IN ACO(K),AC1(K).
C WHEN RESIDU = .FALSE.,TO COMPUTE THE VECTOR XK=XK+BK(RK).
C      THE TWO CONSTANT COEFFICIENTS OF CK ARE STORED IN ACO(K),AC1(K).

      SUBROUTINE RESINV(XR,RX,Y,LK,M,N,DIMXR,DIMY,DIMLK,DIMM,ACO,AC1,
     *                  SQNORM,RESIDU)
      INTEGER DIMXR,DIMY,DIMLK,DIMM,LK(DIMLK),M(DIMM),N(DIMLK)
      REAL XR(DIMXR),RX(DIMXR),Y(DIMY),ACO(DIMLK),AC1(DIMLK)
      LOGICAL SQNORM,RESIDU
      COMMON L,K,SQNM
      SMALL=1.E-35
      IK1=LK(K)
      DO 100 I1=3,IK1
      I3=N(K)+M(I1)
      IK2=I1-1
      DO 100 I2=2,IK2
      I=I3+I2
      YX=0.
      IF(RESIDU) GO TO 77
C IF K=2 AND L NOT EQUAL TO 2, TO COMPUTE X2 = C2(R2).
      IF (K.EQ.2.AND.L.NE.2) GO TO 70
   76 YX=RX(I)
      GO TO 70
C TO COMPUTE RK = RK-AK(XK). IF K=L, R = Y-A(X).
   77 IF(K.NE.L) GO TO 76
      YX=Y(I)
   70 RX(I)=YX+ACO(K)*XR(I)+AC1(K)*(XR(I-1)+XR(I+1)+XR(I-I1)+XR(I-I1+1)+
     *                               XR(I+I1)+XR(I+I1+1))
C TO COMPUTE THE NORM IF REQUIRED.
      IF(SQNORM.AND.ABS(RX(I)).GT.SMALL) SQNM=SQNM+RX(I)**2
  100 CONTINUE
      RETURN
      END
```

Fig. A2

```
C A SUBROUTINE TO SOLVE THE LINEAR SYSTEM A.X = Y IN A TRIANGULAR DOMAIN.
C LK : AN INTEGER ARRAY OF DIMENSION = K, LK(K) = 2**K.
C N : AN INTEGER ARRAY OF DIMENSICN = K; STRUCTURE CONSTANTS, N(K) =
C      TOTAL NUMBER OF POINTS IN THE TRIANGULAR LATTICE IN LEVEL K-1,
C      LEVEL K,...,LEVEL L.
C M : AN INTEGER ARRAY OF DIMENSICN = 1+2**K;STRUCTURE CONSTANTS,M(I1) =
C      TOTAL NUMBER OF POINTS IN ROW1, ROW2,...,ROW(I-1).
C X : AN ARRAY OF DIMENSION = DIMXY, TO STORE THE VECTOR XK, FROM K = L
C      TO K = 2. XK(I1,I2)=X(N(K)+M(I1)+I2).
C R : AN ARRAY OF DIMENSION = DIMXR, TO STORE THE RESIDUAL VECTORS RK,
C      FROM TOP LEVEL K=L TO BOTTOM LEVEL K=2. RK(I1,I2)=R(N(K)+M(I1)+I2)
C IT : ON RETURN,IT SHOWS THE NUMBER OF NORMS COMPUTED.
C NORM : AN INTEGER ARRAY OF DIMENSION = NIT,IT SHOWS THE HISTROY OF THE
C      NORM OF THE RESIDUAL R.
C TOL : SUBROUTINE RETURNS X WHEN NORM OF R HAS LESS THAN THE TOLERANCE
C      TOL OR NUMBER OF ITERATIONS REACHES NIT-1.

      SUBROUTINE FAPIN (X,R,Y,NORM,LK,M,N,DIMXR,DIMY,DIMLK,DIMM,IT,NIT,
     *                   TOL,A0,A1,C0,C1)
      INTEGER DIMXR,DIMY,DIMLK,DIMM,LK(DIMLK),M(DIMM),N(DIMLK)
      REAL NORM(NIT),X(DIMXR),R(DIMXR),Y(DIMY)
      REAL A0(DIMLK),A1(DIMLK),C0(DIMLK),C1(DIMLK)
      COMMON L,K,SQNM
      L=DIMLK
      L1=L-1
      NL=2**L
C TO STORE THE TWO CONSTANT COEFFICIENTS OF -AK IN A0(K),A1(K).
      DO 100 I=2,L
      A0(I)=-A0(I)
      A1(I)=-A1(I)
  100 CONTINUE
C NL2 IS THE TOTAL NUMBER OF INTERIOR POINTS
      NL2=(NL-1)*(NL-2)/2
      NIT1=NIT-1
      DO 901 IT=1,NIT1
      K=L
      SQNM = 0.0
C TO COMPUTE R=Y-A.X.
      CALL RESINV(X,R,Y,LK,M,N,DIMXR,DIMY,DIMLK,DIMM,A0,A1,.TRUE.,
     *            .TRUE.)
      SQNM=SQRT(SQNM/NL2)
      NORM(IT)=SQNM
      IF(SQNM .LT. TOL) RETURN
C IF L = 2 ,TO COMPUTE X2=C(R2).
      IF(L.EQ.2) GO TO 500
C PROJECT RK TO LEVEL K-1.
      DO 800 LL=2,L1
      K=L-LL+1
      JK1=LK(K)
      DO 800 I1=3,JK1
      J3=N(K)+M(I1)
      I3=N(K+1)+M(2*I1-1)
      IK1=I1-1
      DO 800 I2=2,IK1
      J=J3+I2
      I=I3+2*I2-1
  800 R(J)=R(I)+0.5*(R(I-1)+R(I+1)+R(I-2*I1+1)+R(I-2*I1+2)+R(I+2*I1)+
     *                R(I+2*I1-1))
```

```
C  TO COMPUTE  X2 = C(R2).
      CALL RESINV(R,X,Y,LK,M,N,DIMXR,DIMY,DIMLK,DIMM,C0,C1,.FALSE.,
     *               .FALSE.)
C  TO INTERPOLATE X IN THE SPACE K+1 FROM SPACE K.
      K=3
C  TO COMPUTE  XK = QK(XK1),WHERE K1 = K-1.
  600 JK1=LK(K-1)
      DO 300  I1=2,JK1
      J3=N(K-1)+M(I1)
      I3=N(K)+M(2*I1-1)
      DO 300  I2=2,I1
      I=I3+2*I2-1
      J=J3+I2
      IF(K.EQ.L) GO TO 350
      X(I)=X(J)
      X(I-1)=0.5*(X(J)+X(J-1))
      X(I+2*I1-2)=0.5*(X(J-1)+X(J+I1))
      X(I+2*I1-1)=0.5*(X(J)+X(J+I1))
      GO TO 300
C  AT TOP LEVEL L, XL= XL + QL(XL1).
  350 X(I)=X(J)+X(I)
      X(I-1)=0.5*(X(J)+X(J-1))+X(I-1)
      X(I+2*I1-2)=0.5*(X(J-1)+X(J+I1))+X(I+2*I1-2)
      X(I+2*I1-1)=0.5*(X(J)+X(J+I1))+X(I+2*I1-1)
  300 CONTINUE
C  TO COMPUTE  RK = RK-AK(XK).
      CALL RESINV(X,R,Y,LK,M,N,DIMXR,DIMY,DIMLK,DIMM,A0,A1,.FALSE.,
     *               .TRUE.)
C  TO COMPUTE  XK = XK + CK(RK).
  500 CALL RESINV(R,X,Y,LK,M,N,DIMXR,DIMY,DIMLK,DIMM,C0,C1,.FALSE.,
     *               .FALSE.)
      IF(K.EQ.L) GO TO 901
      K=K+1
      GO TO 600
901   CONTINUE
      IT=NIT
C  TO COMPUTE  R = Y-A(X) AND THE NORM OF THE RESIDUAL R.
      SQNM=0.
      CALL RESINV(X,R,Y,LK,M,N,DIMXR,DIMY,DIMLK,DIMM,A0,A1,.TRUE.,
     *               .TRUE.)
      NORM(NIT)=SQRT(SQNM/NL2)
      RETURN
      END
```

APPENDIX B

FORTRAN PROGRAMS FOR PREDICTING THE LIMIT OF SEQUENCE

In Chapter 4, we have mentioned that the convergence of a sequence can sometimes be accelerated by the application of a family of non-linear sequence-to-sequence transformations proposed by D. Shanks [S3]. These transformations are defined as follows. Let $\{x_n\}$ be a sequence of numbers, let

$$\Delta x_n = x_{n+1} - x_n$$

and let k be a positive integer. Then a new sequence $\{B_{k,m}\}$ $(m=k,k+1,k+2,\cdots)$, "the k'th order transform of $\{x_n\}$", is defined, if the denominator does not vanish, by

$$B_{k,m} = \frac{\begin{vmatrix} x_{m-k} & \cdots & x_{m-1} & x_m \\ \Delta x_{m-k} & \cdots & \Delta x_{m-1} & \Delta x_m \\ \Delta x_{m-k+1} & \cdots & \Delta x_m & \Delta x_{m+1} \\ \vdots & & \vdots & \vdots \\ \Delta x_{m-1} & \cdots & & \Delta x_{m+k-1} \end{vmatrix}}{\begin{vmatrix} 1 & \cdots & 1 & 1 \\ \Delta x_{m-k} & \cdots & \Delta x_{m-1} & \Delta x_m \\ \Delta x_{m-k+1} & \cdots & \Delta x_m & \Delta x_{m+1} \\ \vdots & & \vdots & \vdots \\ \Delta x_{m-1} & \cdots & & \Delta x_{m+k-1} \end{vmatrix}} \qquad (1)$$

We observe that the expression in (1) can be written as

$$B_{k,m} = \frac{\begin{vmatrix} \Delta x_{m-1} & \cdot & \Delta x_{m-k+1} & \Delta x_{m-k} & x_{m-k} \\ \vdots & & \vdots & \vdots & \vdots \\ \Delta x_{m+k-2} & \cdots & \Delta x_{m} & \Delta x_{m-1} & x_{m-1} \\ \Delta x_{m+k-1} & \cdots & \Delta x_{m+1} & \Delta x_{m} & x_{m} \end{vmatrix}}{\begin{vmatrix} \Delta x_{m-1} & \cdots & \Delta x_{m-k+1} & \Delta x_{m-k} & 1 \\ \vdots & & \vdots & & \vdots \\ \Delta x_{m+k-2} & \cdots & \Delta x_{m} & \Delta x_{m-1} & 1 \\ \Delta x_{m+k-1} & \cdots & \Delta x_{m+1} & \Delta x_{m} & 1 \end{vmatrix}}$$

and the value $B_{k,m}$ is the solution of the following system of linear equations :

$$\begin{pmatrix} \Delta x_{m-1} & \cdot & \Delta x_{m-k+1} & \Delta x_{m-k} & 1 \\ \vdots & & \vdots & \vdots & \vdots \\ \Delta x_{m+k-2} & \cdots & \Delta x_{m} & \Delta x_{m-1} & 1 \\ \Delta x_{m+k-1} & \cdots & \Delta x_{m+1} & \Delta x_{m} & 1 \end{pmatrix} \begin{pmatrix} Z_0 \\ \\ Z_{k-1} \\ B_{k,m} \end{pmatrix} = \begin{pmatrix} x_{m-k} \\ \\ x_{m-1} \\ x_{m} \end{pmatrix} \qquad (2)$$

Thus the value of $B_{k,m}$ can be obtained by Gaussian Elimination. The whole procedure is carried out by the two subroutines SEQSMT and DETERM as shown in Fig. B1 and Fig. B2 respectively. At the end of the execution, the program SEQSMT returns the transformed sequence $\{B_{k,m}\}$ stored in the array BK and the order of transformation for each term $B_{k,m}$ stored in the integer array ORDER to the calling program.

Fig. B1

```
C SEQSMT IS A SUBROUTINE TO GENERATE A NEW SEQUENCE BK(M) IN ACCELERATING
C THE CONVERGENCE OF SLOWLY CONVERGENT SEQUENCES AND IN INDUCING
C CONVERGENT OF SOME DIVERGENCE SEQUENCES. IN CASE THE MATRIX INDUCE BY
C THE REQUIRED ORDER OF TRANSFORMATION IS SINGULAR, THE ORDER OF
C       TRANSFORMATION WILL BE REDUCED TO A LOWER ORDER.
C PARAMETERS OF THE SUBROUTINE REQUIRE:
C       1. X: AN ARRAY OF THE ORIGINAL SEQUENCE
C       2. N: DIMENSION OF THE ARRAY X
C       3. K: THE ORDER OF TRANSFORMATION OF THE SEQUENCE X(N) (BETWEEN 0
C                                                         AND (N-1)/2).
C       4. BK: REAL ARRAY, TO STORE THE GENERATED NEW SEQUENCE.
C       5. DIMBK: DIMENSION OF BK, DIMBK = N-2*K.
C       6. A : AN DUMMY ARRAY OF DIAMENSION KP1 BY KP2
C       7. KP1 : EQUAL TO K+1
C       8. KP2 : EQUAL TO K+2
C       9. ORDER : AN INTEGER ARRAY (DIMENSION=DIMBK) TO STORE THE ORDER
C                  OF TRANSFORMATION.

      SUBROUTINE SEQSMT(X,N,K,BK,DIMBK,A,KP1,KP2,ORDER)
      INTEGER DIMBK,ORDER
      DIMENSION X(N),BK(DIMBK),A(KP1,KP2),ORDER(DIMBK)
      IF(N.GE.2*K+1) GO TO 4
      KK=(N-1)/2
C IF ORDER OF TRANSFORMATION IS OUT OF RANGE,STOP RUN.
      WRITE (6,66) KK
   66 FORMAT ('0','ORDER OF TRANSFORMATION MUST BE BETWEEN 1 AND ', I2)
      STOP
    4 NMK=N-K
      DO 100 MK=KP1,NMK
      K1=KP1
      K2=KP2
      CALL DETERM(X,N,MK,K1,K2,A,BKM,&1)
      GO TO 110
C IF THE COEFFICIENT MATRIX OF THE LINEAR EQUATIONS IS SINGULAR, REDUCE
C     THE ORDER OF TRANSFORMATION FOR THE TERM BK(M) BY 1.
    1 K1=K1-1
      K2=K2-1
      CALL DETERM(X,N,MK,K1,K2,A,BKM,&1)
  110 ORDER(MK-K)=K1-1
  100 BK(MK-K)=BKM
      RETURN
      END
```

```
C THE METHOD OF GAUSSIAN ELIMINATION TO COMPUTE THE RATIO OF TWO
C      DETERMINANTS.
C PARTIAL PIVOTAL CONDENSATION IS USED- A SEARCH IS MADE IN EACH COLUMN
C      FOR THE LARGEST ELEMENT BELOW THE DIAGONAL,BUT OTHER COLUMNS ARE
C      NOT SEARCHED.

      SUBROUTINE DETERM(X,N,MK,KP1,KP2,A,BKM,*)
      DIMENSION A(KP1,KP2),X(N)
      SMALL=0.1E-35
      IF (KP1.GE.2) GO TO 100
      BKM=X(MK)
      RETURN
C TO CREAT THE AUGMENTED MATRIX A(I,J)
  100 K=KP1-1
      DO 750 I=1,KP1
      DO 700 J=1,K
      II=MK+I-J
  700 A(I,J)=X(II)-X(II-1)
      A(I,KP1)=1.
  750 A(I,KP2)=X(MK+I-KP1)
C BEGIN THE PARTIAL PIVOTAL CONDENSATION
      DO 600 II=1,K
      IIP1=II+1
      L=II
C FIND TERM IN COLUMN II,ON OR BELOW MAIN DIAGONAL, THAT IS LARGEST IN
C      ABSOLUTE VALUE. AFTER THE SEARCH, L IS THE ROW NUMBER OF THE
C      LARGEST ELEMENT.
      DO 400 I=IIP1,KP1
  400 IF(ABS(A(I,II)).GT.ABS(A(L,II))) L=I
C IF THE MATRIX IS SINGULAR ,RETURN BACK TO THE CALLING PROGRAM TO
C      REDUCE THE ORDER OF TRANSFORMATION BY 1 AND REENTER THIS SUBPROGRAM
      IF (ABS(A(L,II)).LT.SMALL) RETURN1
      IF (L.EQ.II) GO TO 500
C INTERCHANGE ROWS L AND II, FROM DIAGONAL RIGHT
      DO 410 J=II,KP2
      TEMP=A(II,J)
      A(II,J)=A(L,J)
  410 A(L,J)=TEMP
C ELIMINATE ALL ELEMENTS IN COLUMN II BELOW MAIN DIAGONAL
  500 DO 600 I=IIP1,KP1
      FACTOR=A(I,II)/A(II,II)
      DO 600 J=IIP1,KP2
  600 A(I,J)=A(I,J)-FACTOR*A(II,J)
C IF THE MATRIX IS SINGULAR ,RETURN BACK TO THE CALLING PROGRAM TO
C      REDUCE THE ORDER OF TRANSFORMATION BY 1 AND REENTER THIS SUBPROGRAM
      IF(ABS(A(KP1,KP1)).LT.SMALL) RETURN1
      BKM=A(KP1,KP2)/A(KP1,KP1)
      RETURN
      END
```

Fig. B2

APPENDIX C

FORTRAN PROGRAMS TO COMPUTE THE L^2 norm of the function $U-U^h$

This appendix contains FORTRAN FUNCTION subprograms to compute the L^2 norm of the error functional $U-U^h$, where U^h is the Ritz-Galerkin solutions to U in the finite dimensional subspace $S_g^{1,0}$.

Fig. C2 contains the FUNCTION subprogram BYCO to compute the Barycentric Coordinates of the integer lattice (i_1, i_2) (see Appendix A) w.r.t. the triangle $X_0X_1X_2$.

Fig. C1 contains the FUNCTION subprogram L2SQ. It interpolates the function U^h and then computes the square of the L^2 norm of the function $U-U^h$ in each of the triangle $Y_0Y_1Y_2$ by using some numerical quadratures on a triangle T. The Barycentric Coordinates of the three vertices Y_0, Y_1, Y_2 are given by the calling program, and the Barycentric Coordinates of each point $X(x_0, x_1, x_2)$ in $Y_0Y_1Y_2$ w.r.t. the large triangle $X_0X_1X_2$ are computed according to the linear transformation (1.3.4) given in Chapter 1.

Fig. C3 contains the FUNCTION subprogram L2NORM. It computes the L^2 norm of $U-U^h$ over the triangle $X_0X_1X_2$.

```
C A FUNCTION SUBPROGRAM TO COMPUTE THE SQUARE OF THE L2 NORM OF THE
C      FUNCTION (XINT - U) IN THE TRIANGLE YOY1Y2.
C      XINT IS THE LINEAR INTERPOLATION OF THE FUNCTION W IN THE TRIANGLE
C      YOY1Y2.
C THE BARYCENTRIC  COORDINATES OF THE THREE VERTICES Y0,Y1,Y2 ARE STORED
C      IN THE ARRAY Y0(3),Y1(3) AND Y2(3) RESPECTIVELY.
C W0,W1,W2 ARE THE VALUES OF W AT Y0,Y1,Y2 RESPECTIVELY.
C NH IS THE NUMBER OF INTERVALS TO BE DIVIDED ON EACH SIDE OF THE
C      TRIANGLE YOY1Y2.
C THE QUADRATURE COEFFICIENTS ARE STORED IN THE ARRAY QUADT, THEY ARE
C      NUMBERED FROM TOP TO BOTTOM AND FROM LEFT TO RIGHT.
C NOQUAD : DIMENSION OF QUADT,NOQUAD = (NH+1)*(NH+2)/2.

      FUNCTION L2SQ(NH,NOQUAD,QUADT,W0,W1,W2,Y0,Y1,Y2)
      REAL L2SQ,QUADT(NOQUAD),Y0(3),Y1(3),Y2(3),Z(3)
      SMALL=1.E-35
      H2=1./NH
      L2SQ=0.
      I=0
      NP1=NH+1
      DO 700 I1=1,NP1
      DO 700 I2=1,I1
      I=I+1
C     TO COMPUTE THE LOCAL BARYCENTRIC COORDINATES OF Z W.R.T. THE
C     TRIANGLE YOY1Y2.
      CALL BYCO(I1,I2,H2,Z)
C     TO COMPUTE THE BARYCENTRIC COORDINATES OF Z W.R.T. THE LARGE
C     TRIANGLE T, THE DOMAIN OF U.
      X0=Y0(1)*Z(1)+Y1(1)*Z(2)+Y2(1)*Z(3)
      IF(ABS(X0).LT.SMALL) X0=0.
      X1=Y0(2)*Z(1)+Y1(2)*Z(2)+Y2(2)*Z(3)
      IF(ABS(X1).LT.SMALL) X1=0.
      X2=1.-X0-X1
      IF(ABS(X2).LT.SMALL) X2=0.
      XINT=Z(1)*W0+Z(2)*W1+Z(3)*W2
      DIFF=XINT-U(X0,X1,X2)
      IF(ABS(DIFF).GT.SMALL) L2SQ=L2SQ+QUADT(I)*DIFF**2
  700 CONTINUE
      RETURN
      END
```

Fig. C1

Fig. C2

```
C       TO COMPUTE THE BARYCENTRIC COORDINATES OF A POINT IN THE TRIANGLE
C       T.

        SUBROUTINE BYCO(I1,I2,H,BC)
        REAL BC(3)
        SMALL=1.E-35
        BC(1)=1.-(I1-1.)*H
        IF(ABS(BC(1)).LT.SMALL) BC(1)=0.
        BC(3)=(I2-1.)*H
        IF(ABS(BC(3)).LT.SMALL) BC(3)=0.
        BC(2)=1.-BC(1)-BC(3)
        IF(ABS(BC(2)).LT.SMALL) BC(2)=0.
        RETURN
        END
```

Fig. C3

```
C A FUNCTION SUBPROGRAM TO COMPUTE THE L2 NORM OF THE FUNCTION (U - X)
C       IN A TRIANGULAR DOMAIN T, WHERE X IS THE RITZ-GALERKIN SOLUTION TO
C       U AT LEVEL L.
C QUACON IS THE QUADRATURE NORMALIZE CONSTANT.
C NL = 2**L.

        FUNCTION L2NORM(X,M,DIMX,DIMM,NL,NH,NOQUAD,QUADT,QUACON)
        INTEGER DIMX,DIMM,M(DIMM)
        REAL L2NORM,QUADT(NOQUAD),X(DIMX),Y0(3),Y1(3),Y2(3)
        REAL L2SQ
        ERROR=0.
        H=1./NL
        DO 90 I1=1,NL
        DO 90 I2=1,I1
        I=M(I1)+I2
        CALL BYCO(I1,I2,H,Y0)
        CALL BYCO(I1+1,I2,H,Y1)
        CALL BYCO(I1+1,I2+1,H,Y2)
C Y0,Y1,Y2 ARE THE BARRYCENTRIC COORDINATES OF THE THREE VERTICES OF T.
        ERROR=ERROR+L2SQ(NH,NOQUAD,QUADT,X(I),X(I+I1),X(I+I1+1),Y0,Y1,Y2)
        IF(I2.EQ.I1) GO TO 90
        CALL BYCO(I1,I2+1,H,Y1)
        ERROR=ERROR+L2SQ(NH,NOQUAD,QUADT,X(I),X(I+ 1),X(I+I1+1),Y0,Y1,Y2)
     90 CONTINUE
        L2NORM=SQRT(ERROR*QUACON)
        RETURN
        END
```

APPENDIX D

FORTRAN PROGRAMS TO CONSTRUCT THE DISCRETE EIGENVALUE λ^h

This appendix contains four FORTRAN subprograms to solve the generalized eigenvalue problem

$$A^h x^h = \lambda^h B^h x^h$$

in a triangular domain Ω.

The algorithm can be described as [F7]

$$\begin{aligned}
r^{(k)} & \quad (A^h - \lambda^h B^h) x^{(k)} \\
w^{(k)} & \quad B^h x^{(k)} \\
\nu^{(k)} & \quad (r^{(k)}, x^{(k)})/(w^{(k)}, x^{(k)}) \\
x^{(k+1)} & \quad x^{(k)} - C^h r^{(k)} \\
\lambda^{(k+1)} & \quad \lambda^{(k)} + \nu^{(k)}
\end{aligned} \qquad (1)$$

Fig. D1 contains the subprogram RESIDU. It computes the residual $r^{(k)}$, the vector $w^{(k)}$ and the approximate eigenvector $x^{(k)}$. The inner products $(r^{(k)},x^{(k)})$ and $(w^{(k)},x^{(k)})$ are also computed in this subprogram while evaluating $r^{(k)}$ and $w^{(k)}$ by setting INNPRO = .TRUE.

Fig. D2 contains the subprogram APRINV. It constructs an WT_q ε-approximate inverse C^h to $A^h - \lambda^{(k)} B^h$ by calling the subprogram Gauss listed in Fig. D3 to solve a system of linear equations.

The step (1) is not executed unless $|\nu^{(k)}-\nu^{(k-1)}| < EPS$, where EPS is a given constant. After $|\nu^{(k)}-\nu^{(k-1)}|$ is less than the tolerance TOL or the number of iterations reaches NIT, the subprogram EIGEN returns a series of sucessive approximate eigenvalues to the calling program.

Fig. D1

```
C-----TO REFINE THE EIGENVECTOR X AND COMPUTE THE RESIDUAL=(A-EIGVAL.I)X
C-----TO COMPUTE THE VECTOR RK OR WK, ACO=AA0, AC1=AA1, XR=X, RX=R,
C                                  INNPRO=.TRUE.
C-----TO REFINE THE EIGENVECTOR X, ACO=-C0, AC1=-C1, XR=R, RX=X,
C                                  INNPRO=.FALSE.

      SUBROUTINE RESIDU(XR,RX,M,DIMXR,DIMM,ACO,AC1,SUMRX,NL,INNPRO)
      INTEGER DIMXR,DIMM,M(DIMM)
      REAL XR(DIMXR),RX(DIMXR)
      LOGICAL INNPRO
      SMALL=1.E-35
      DO 100 I1=3,NL
      IK1=I1-1
      DO 100 I2=2,IK1
      I=M(I1)+I2
      YX=RX(I)
      IF(INNPRO) YX=0.
      RX(I)=YX+ACO*XR(I)+AC1*(XR(I-1)+XR(I+1)+XR(I-I1)+XR(I-I1+1)+
     *                        XR(I+I1)+XR(I+I1+1))
C-----TO COMPUTE THE INNER PRODUCT IF INNPRO = .TRUE.
      IF(INNPRO.AND.ABS(RX(I)).GT.SMALL.AND.ABS(XR(I)).GT.SMALL)
     *   SUMRX=SUMRX+XR(I)*RX(I)
  100 CONTINUE
      RETURN
      END
```

```
C-----CONSTRUCTION OF THE WTQ APPROXIMATION INVERSE C OF THE LINEAR
C     OPERATOR A WITH WEIGHTS W0,W1.

      SUBROUTINE APRINV(A0,A1,C0,C1,W0,W1)
      REAL A(2,3),C(2)
      A(1,3)=W0
      A(2,3)=W1
      A(1,1)=A0*W0+6.*A1*W1
      A0P2A1=A0+2.*A1
      A(2,1)=A1*W0+A0P2A1*W1
      A(1,2)=6.*A(2,1)
      A(2,2)=A0P2A1*W0+(15.*A1+2.*A0)*W1
      CALL GAUSS (A,C,2,3,&1)
      C0=C(1)
      C1=C(2)
      RETURN
    1 WRITE(6,77)
   77 FORMAT('    THE AUGMENTED MATRIX IS SINGULAR')
      STOP
      END
```

Fig. D2

```
C THE METHOD OF GAUSSIAN ELIMINATION FOR SOLVING SIMULTANEOUS LINEAR
C      EQUATIONS.
C PARTIAL PIVOTAL CONDENSATION IS USED- A SEARCH IS MADE IN EACH COLUMN
C      FOR THE LARGEST ELEMENT BELOW THE DIAGONAL,BUT OTHER COLUMNS ARE
C      NOT SEARCHED.

      SUBROUTINE GAUSS (A,X,N,NP1,*)
      REAL A(N,NP1),X(N)
      SMALL=0.1E-35
      NM1=N-1
C BEGIN THE PARTIAL PIVOTAL CONDENSATION
      DO 600 K=1,NM1
      KP1=K+1
      L=K
C FIND TERM IN COLUMN K, ON OR BELOW MAIN DIAGONAL, THAT IS LARGEST IN
C      ABSOLUTE VALUE. AFTER THE SEARCH, L IS THE ROW NUMBER OF THE
C      LARGEST ELEMENT.
      DO 400 I=KP1,N
  400 IF(ABS(A(I,K)).GT.ABS(A(L,K))) L=I
      IF (ABS(A(L,K)).LE.SMALL) RETURN1
      IF (L.EQ.K) GO TO 500
C INTERCHANGE ROWS L AND K, FROM DIAGONAL RIGHT
      DO 410 J=K,NP1
      TEMP=A(K,J)
      A(K,J)=A(L,J)
  410 A(L,J)=TEMP
C ELIMINATE ALL ELEMENTS IN COLUMN K BELOW MAIN DIAGONAL
  500 DO 600 I=KP1,N
      FACTOR=A(I,K)/A(K,K)
      DO 600 J=KP1,NP1
  600 A(I,J)=A(I,J)-FACTOR*A(K,J)
C BACK SUBSTITUTION
      IF(ABS(A(N,N)).LT.SMALL) RETURN1
      X(N)=A(N,NP1)/A(N,N)
      DO 710 IN=1,NM1
      I=N-IN
      IP1=I+1
      SUM=0.
      DO 700 J=IP1,N
  700 SUM=SUM+A(I,J)*X(J)
  710 X(I)=(A(I,NP1)-SUM)/A(I,I)
      RETURN
      END
```

Fig. D3

```
C-----TO SOLVE THE GENERALIZED EIGENVALUE PROBLEM : A.X = EIGVAL.B.X
C-----W0, W1 ARE THE TWO CONSTANT COEFFICIENTS OF THE WEIGHT W.
C-----PROGRAM RETURNS THE APPROXIMATE EIGENVALUES, IF THE DIFFERENCE
C      BETWEEN TWO SUCCESSIVE EIGENVALUES LESS THAN TOL OR NUMBER OF
C      ITERATIONS REACHES NIT-1.
C-----X IS AN APPROXIMATE EIGENVECTOR OF A, THE INITIAL APPROXIMATION
C      CAN BE RANDOM.
C-----EIGVAL : AN ARRAY CONTAINS THE SUCCESSIVE APPROXIMATE EIGENVALUES.
C-----ON RETURN,IT SHOWS THE NUMBER OF RECORDED SUCCESSIVE APPROXIMATE
C      EIGENVALUES.
C-----M : STRUCTURE CONSTANTS.
C-----EPS : A CONSTANT, IF DIFFERENCE BETWEEN TWO SUCCESSIVE RATIO .GT.
C      EPS, RATIO IS NOT ADDED TO EIGVAL(IT).
C-----A0,A1 ARE THE TWO CONSTANT COEFFICIENTS OF THE LAPLACIAN OPERATOR.
C-----NL = 2**L.

      SUBROUTINE EIGEN(X,R,EIGVAL,M,DIMXR,DIMM,IT,NIT,EPS,A0,A1,TOL,NL,
     *                 W0,W1)
      INTEGER DIMXR,DIMM,M(DIMM)
      REAL X(DIMXR),R(DIMXR),EIGVAL(NIT)
      RATIO0=0.
      SMALL=1.E-35
      HH=1./NL**2
      H0=0.75*HH
      H1=0.125*HH
      IT=1
      DO 999 ITER=1,NIT
      IF(ABS(EIGVAL(IT)).LE.SMALL) GO TO 33
C-----SET UP THE TWO CONSTANT COEFFICIENTS OF THE OPERATOR (A-EIGVAL.B).
      AA0=A0-H0*EIGVAL(IT)
      AA1=A1-H1*EIGVAL(IT)
C-----TO COMPUTE THE RESIDUAL RK AND THE INNER PRODUCT (RK,XK).
   33 SUMRX=0.
      CALL RESIDU(X,R,M,DIMXR,DIMM,AA0,AA1,SUMRX,NL,.TRUE.)
C-----CALL THE APRINV SUBROUTINE TO CONSTRUCT AN APPROXIMATE INVERSE OF
C      A-EIGVAL.B
      CALL APRINV(AA0,AA1,C0,C1,W0,W1)
      CALL RESIDU(R,X,M,DIMXR,DIMM,-C0,-C1,SUMWX,NL,.FALSE.)
C-----TO COMPUTE THE VECTOR WK AND THE INNER PRODUCT  (WK,XK).
      AA1=1.
      AA0=6.
      SUMWX=0.
      CALL RESIDU(X,R,M,DIMXR,DIMM,AA0,AA1,SUMWX,NL,.TRUE.)
      SUMWX=SUMWX*H1
      IF(ABS(SUMWX).LT.SMALL) GO TO 999
      RATIO=SUMRX/SUMWX
      IF(ABS(RATIO-RATIO0).GT.EPS) GO TO 999
      RATIO0=RATIO
      EIGVAL(IT+1)=EIGVAL(IT)+RATIO
      IF(ABS(RATIO).LT.TOL) RETURN
      IT=IT+1
  999 CONTINUE
      RETURN
      END
```

Fig. D4

APPENIX E

FORTRAN PROGRAMS TO SOLVE THE POISSON EQUATION $LU = -\Delta U + \lambda U = f$ IN A TRIANGULAR DOMAIN Ω

This appendix contains FORTRAN programs to solve the boundary value problem

$$\begin{cases} LU = -\Delta U + \lambda U = f & \text{in } \Omega \\ U = g & \partial\Omega \end{cases}$$

Fig. E1 is the Fortran subroutine SPRANY, to produce an analysis report of the norms of the residue r and the spectral radius of the linear operator $I-C^hL^h$, where C^h is an ε-approximate inverse to the discrete linear operator L^h.

Fig. E2 contains the FORTRAN subroutine PRINTG to print out the vector X, Y or R in an triangular form.

Fig. E3 contains the FUNCTION subprogram U, the exact solution of $LU = f$.

Fig. E4 contains the main program to construct the Ritz-Galerkin solution to $LU = f$.

Fig. E1

```
C-----ANALYSIS OF NORM AND SPECTRAL RADIUS.
C-----THIS SUBROUTINE CALLS THE SEQSMT SUBROUTINE TO ACCELERATE THE
C     CONVERGENCE OF THE SEQUENCE OF SPECTRAL RADIUS, AND OUTPUT A
C     LISTING OF THE ANALYTICAL RESULTS.
C-----K IS THE ORDER OF TRANSFORMATION.
C-----A IS A KP1 BY KP2 DUMMY ARRAY, WHERE KP1 = K+1, KP2 = K+2.
C-----SPECTR : A IT BY 5 REAL ARRAY TO STORE THE SPECTRAL RADIUS AND
C     THE SMOOTHED SPECTRAL RADIUS.
C-----ORDER : A IT BY 4 INTEGER ARRAY TO STORE THE ORDER OF
C     TRANSFORMATION OF THE SMOOTHED SPECTRAL RADIUS.

      SUBROUTINE SPRANY(NORM,SPECTR,ORDER,K,IT,A,KP1,KP2)
      INTEGER ORDER(IT,4),IORD(4)
      REAL NORM(IT),SPECTR(IT,5),P1(2),P2(4),A(KP1,KP2)
      IT1=IT-1
C-----TO GENERATE A SEQUENCE OF SPECTRAL RADIUS.
      DO 44 I=1,IT1
   44 SPECTR(I,1)=NORM(I+1)/NORM(I)
   83 KP1=K+1
      KP2=K+2
      DO 41 I=1,4
      IT3=IT1-2*K*I
      IT2=IT3+2*K
      IF(IT3.GT.0) GO TO 1
      IF(K.LT.1.OR.I.GT.1) GO TO 39
C-----TO REDUCE THE ORDER OF TRANSFORMATION BY 1, IF THE NUMBER OF TERMS
C     IN THE SEQUENCE ARE NOT ENOUGH TO CARRY OUT THE REQUIRED ORDER OF
C     TRANSFORMATION.
      K=K-1
      GO TO 83
C-----CALL THE SEQSMT SUBROUTINE TO PERFORM A NON-LINEAR TRANSFORMATION.
    1 CALL SEQSMT(SPECTR(1,I),IT2,K,SPECTR(1,I+1),IT3,A,KP1,KP2,
     *             ORDER(1,I))
   41 CONTINUE
   87 FORMAT ('-','ITERAT  NORM               NORM(I)/NORM(I-1)  TRANSFORMA
     *TION ORDER/SMOOTHED SPECTRAL RADIUS',/,' ',35X,4(8X,'ITERATION ',
     *I1))
   39 ITER=1
      IF(K.EQ.0) GO TO 88
C-----TO DETERMINE HOW MANY TIME OF ITERATIVE TRANSFORMATIONS HAS BEEN
C     PERFORMED.
      ITER=(IT-2)/(2*K)
      IF(ITER.GT.4) ITER=4
   88 WRITE(6,87) (I,I=1,ITER)
C-----OUTPUT THE ANALYTICAL RESULTS.
      II2=0
      DO 80 I=1,IT
      I1=I-1
      II=1
      IF(I  .EQ. 1) GO TO 10
      II=2
      II2=0
      DO 85 I2=1,4
      J1=K*I2
      IF(I1.LE.J1.OR.IT-I1.LE.J1) GO TO 20
      P2(I2)=SPECTR(I1-J1,I2+1)
      IORD(I2)=ORDER(I1-J1,I2)
      II2=I2
   85 CONTINUE
   20 P1(2)=SPECTR(I1,1)
   10 P1(1)=NORM(I)
      IF(II2 .EQ. 0) GO TO 65
      WRITE(6,70) I1,(P1(J),J=1,II), (IORD(J1),P2(J1),J1=1,II2)
   70 FORMAT(2X,I2,4X,E14.7,2X,E14.7,4X,4(1X,I2,2X,E14.7))
      GO TO 80
   65 WRITE(6,70) I1,(P1(J),J=1,II)
   80 CONTINUE
      RETURN
      END
```

Fig. E2

```
C-----PRINT OUT THE CONTENTS OF VALUES IN THE ARRAY X,R OR Y IN A
C      TRIANGULAR FORM.
C-----K SPECIFY THE LEVEL AT WHICH X OR R TO BE PRINTED.
C-----TO PRINT X IF XYR = 1.
C-----TO PRINT Y IF XYR = 2.
C-----TO PRINT R IF XYR = 3.

      SUBROUTINE PRINTG(X, R, Y,LK,M,N,DIMXR,DIMY,DIMLK,DIMM,XYR,K)
      INTEGER DIMXR,DIMY,DIMLK,DIMM,XYR,LK(DIMLK),M(DIMM),N(DIMLK)
      REAL X(DIMXR),R(DIMXR),Y(DIMY),P(8)
      J=1
      IK0=1
      IK1=LK(K)+1
      IF(XYR.NE.2) GO TO 70
C-----TO PRINT THE INTERIOR POINTS OF Y.
      IK0=3
      IK1=LK(K)
   70 DO 200 I1=IK0,IK1
      J3=N(K)+M(I1)
      IK3=1
      IK4=I1
      IF(XYR.NE.2) GO TO 72
      IK3=2
      IK4=I1-1
   72 DO 250 I2=IK3,IK4
C-----ONLY PRINT OUT THE FIRST 8 VALUES IN EACH ROW.
      IF(J.GT.8) GO TO 24
      I=J3+I2
      GO TO (15,16,17),XYR
   15 P(J)=X(I)
      GO TO 25
   16 P(J)=Y(I)
      GO TO 25
   17 P(J)=R(I)
   25 J=J+1
  250 CONTINUE
   24 I3=J-1
   66 FORMAT (8E16.7)
      WRITE (6,66) (P(J),J=1,I3)
      J=1
  200 CONTINUE
      J=J-1
      IF(J.GT.0) WRITE (6,66) (P(I),I=1,J)
      RETURN
      END
```

Fig. E3

```
      FUNCTION U(X0,X1,X2)
      U=SIN(X0)*SIN(X1)*SIN(X0-X1)
      RETURN
      END
```

Fig. E4

```
C-----TO SOLVE THE LINEAR SYSTEM A.X = Y WITH THE HOMOGENEOUS BOUNDARY
C      CONDITICN,THOSE BOUNDARY VALUES OF X AT EACH LEVEL K MUST BE
C      ZEROIZED.
C-----WITH THE INHOMOGENEOUS BOUNDARY CONDITION X = G,THE BOUNDARY VALUES
C      OF X AT THE TOP LEVEL L EQUAL TO THE CORRESPONDING VALUES OF G,AND
C      ALL THE BOUNDARY VALUES OF X AT THE OTHER LEVEL K ARE SET TO ZERO.

      INTEGER ORDER(40,4)
      INTEGER LK(5),M(33),N(5)
      INTEGER DIMXR,DIMY,DIMM
      REAL NORM(40),X(774),R(774),Y(561),A(5,6),L2ERR(5),RATE(4)
      REAL AA0(5),AA1(5),CC0(5),CC1(5),QUADT(6),L2NORM,XX(3)
      DIMENSICN SPECTR(40,5)
      LOGICAL ONEPT,SEVENP
      EQUIVALENCE (XX(1),X0),(XX(2),X1),(XX(3),X2)
      F(X0,X1,X2)=3.*SIN(2.*(X0-X1))-SIN(2.*X0)+SIN(2.*X1)-
     *            39.47842*SIN(XC)*SIN(X1)*SIN(X0-X1)
      SMALL=1.E-35
C-----IF THE 1-POINT NUMERICAL QUADRATURE IS USED, ONEPT=.TRUE.
      CNEPT=.TRUE.
C-----IF THE 7-POINT NUMERICAL QUADRATURE IS USED, SEVENP=.TRUE.
      SEVENP=.TRUE.
      DIMM=33
      DIMY=561
      DIMXR=774
C-----A0,A1 ARE THE TWO CONSTANT COEFFICIENTS OF THE DISCRETE LINEAR
C      OPERATOR A.
      A0=6.
      A1=-1.
C-----SET UP THE QUADRATURE COEFFICIENTS FOR COMPUTING THE L2 NORM.
      NH=2
      NOQUAD=6
      QUADT(1)=0.
      QUADT(2)=1.
      QUADT(3)=1.
      QUADT(4)=0.
      QUADT(5)=1.
      QUADT(6)=0.
      IF(.NOT.ONEPT) GO TO 77
   10 DO 112 L=2,5
      L1=L-1
      K=L
      NL=2**L
      H=1./NL
      HH=H**2
      QUACON=HH/3.
      EIGV=-52.81
C-----TO COMPUTE THE TWO CONSTANT COEFFICIENTS OF AK AND CK.
      BK=EIGV*0.125*HH
      DO 600 I=1,L1
      K=L-I+1
      AA0(K)= 6.*(1.+BK)
      AA1(K)=BK-1.
C-----TO COMPUTE THE APPROXIMATE INVERSE CK.
      CALL APRINV (AA0(K),AA1(K),CC0(K),CC1(K))
      BK=4.*BK
  600 CONTINUE
```

```
   82 FORMAT('-   K',9X,'AA0',18X,'AA1',18X,'CC0',18X,'CC1')
      WRITE(6,82)
   85 FORMAT('   ',I2,1X,4(5X,E16.7))
      DO 74 J=2,L
      WRITE(6,85) J,AA0(J),AA1(J),CC0(J),CC1(J)
   74 CONTINUE
C-----CONSTRUCT THE STRUCTURE CONSTANTS, M.
      M(1)=0
      IK1=NL+1
      DO 30 I=2,IK1
      I1=I-1
   30 M(I)=M(I1)+I1
C-----CONSTRUCT THE STRUCTURE CONSTANTS, LK.
      LK(1)=2
      DO 50 I=2,L
   50 LK(I)=2*LK(I-1)
C-----CONSTRUCT THE STRUCTURE CONSTANTS, N.
      N(L)=0
      DO 40 I=1,L1
      K1=L-I
   40 N(K1)=N(K1+1)+(1+LK(K1))*(1+LK(K1+1))
      IK1=N(1)
      DO 777 I=1,IK1
      X(I)=0.
  777 R(I)=0.
      SQNM=0.
      IK1=NL+1
C-----APPLY THE 7-POINT OR 1-POINT FORMULA TO SET UP THE VECTOR Y.
      DO 768 I1=1,IK1
      DO 768 I2=1,I1
      I=I2+M(I1)
      CALL BYCO(I1,I2,H,XX)
C-----PRESET THE BOUNDARY VALUES OF X AT THE TOP LEVEL L.
      IF(I1.EQ.IK1.OR.I2.EQ.1.OR.I2.EQ.I1) X(I)=U(X0,X1,X2)
      IF(ONEPT) GO TO 768
      R(I)=F(X0,X1,X2)
  768 CONTINUE
      F0=1.5*HH
      F1=HH/12.
  800 H2=2.*HH
      DO 100 I1=3,NL
      IK2=I1-1
      DO 100 I2=2,IK2
      I=I2+M(I1)
      IF(ONEPT) GO TO 405
      Y(I)=F0*R(I)+F1*(R(I-1)+R(I+1)+R(I-I1)+R(I-I1+1)+R(I+I1)+
     *                 R(I+I1+1))
      GO TO 170
  405 CALL BYCO(I1,I2,H,XX)
      Y(I)=H2*F(X0,X1,X2)
  170 IF(ABS(Y(I)).GT.SMALL) SQNM=SQNM+Y(I)**2
  100 CONTINUE
      EPS=1.E-08
C-----COMPUTE THE NORM OF Y.
      NL2=(NL-1)*(NL-2)/2
      SQNM=SQRT(SQNM/NL2)
      IF(SQNM.GT.SMALL) EPS=EPS*SQNM
      NIT=40
C-----ZEROIZED THE VECTOR R.
      IK1=N(L1)
      DO 750 I=1,IK1
  750 R(I)=0.
```

```
C------CALL FAPIN TO SOLVE THE LINEAR SYSTEM A.X = Y.
      CALL FAPIN (X,R,Y,NORM,LK,M,N,DIMXR,DIMY,L,DIMM,IT,NIT,EPS,AA0,AA1
     *           ,CC0,CC1)
C------ENTER THE SPRANY SUBROUTINE TO ANALYSE THE NORM AND SPECTRAL RADIUS
C      AND OUTPUT THE ANALYTICAL RESULTS.
      K=4
      KP1=K+1
      KP2=K+2
      CALL SPRANY(NORM,SPECTR,ORDER,K,IT,A,KP1,KP2)
C------PRINT OUT THE RITZ-GALERKIN SOLUTIONS.
      WRITE(6,122)
  122 FORMAT('-   RITZ-GALERKIN SOLUTIONS, X =')
      CALL PRINTG(X,R,Y,LK,M,N,DIMXR,DIMY,L,DIMM,1,L)
C------PRINT OUT THE INTERIOR POINTS OF Y.
  121 FORMAT('-    INTERIOR POINTS OF Y =')
      WRITE(6,121)
      CALL PRINTG(X,R,Y,LK,M,N,DIMXR,DIMY,L,DIMM,2,L)
      IK1=NL+1
C------PRINT OUT THE EXACT SOLUTIONS OF X.
      DO 867 I1=1,IK1
      DO 867 I2=1,I1
      CALL BYCO(I1,I2,H,XX)
      I=I2+M(I1)
      R(I)=U(X0,X1,X2)
  867 CONTINUE
  123 FORMAT('-    EXACT SOLUTIONS, X =')
      WRITE(6,123)
      CALL PRINTG(X,R,Y,LK,M,N,DIMXR,DIMY,L,DIMM,3,L)
C------ERROR ANALYSIS : COMPUTATION OF L2 NORM AND RATE OF CONVERGENCE.
      L2ERR(L)=L2NORM(X,M,DIMY,DIMM,NL,NH,NCQUAD,QUADT,QUACON)
    6 FORMAT('  L2ERR= ',E16.7)
      WRITE (6,6) L2ERR(L)
      IF(L.GT.2) RATE(L-2)=(ALOG(L2ERR(L1))-ALOG(L2ERR(L)))/ALOG(2.)
  112 CONTINUE
C------PRINT OUT THE ANALYTICAL RESULTS.
   61 FORMAT('-','LEVEL    H',18X,'L2 NORM',12X,'CONVERGENCE RATE')
      WRITE(6,61)
   63 FORMAT(3X,I1,4X,E14.7,2(5X,E14.7))
      L=5
      DO 62 I=2,L
      H=2.**(-I)
      IF(I.EQ.2) GO TO 67
      WRITE(6,63) I,H,L2ERR(I),RATE(I-2)
      GO TO 62
   67 WRITE(6,63) I,H,L2ERR(I)
   62 CONTINUE
   77 IF(.NOT.SEVENP) GO TO 64
      ONEPT=.FALSE.
      SEVENP=.FALSE.
      GO TO 10
   64 STOP
      END
```

BIBLIOGRAPHY

[A1] Aziz, A.K., The Mathematical Foundations of the Finite Element Method with Application to Partial Differential Equations, Academic Press, New York (1972).

[B1] Barnhill, R.E. and Gregory J.A., Interpolation Remainder Theorems on Triangular Domains with Application to Finite Element Error Bounds, Numer. Math. 25, P.215-229 (1976).

[B2] Barnhill, R.E. and Gregory J.A., Interpolation Remainder Theory from Taylor Expansions on Triangles, Numer. Math. P. 401-407 (1976).

[B3] Benson, M.W., Iterative solution of Large Scale Linear Systems, Thesis, Lakehead University (1973).

[B4] Birkhoff, The Numerical Solution of Elliptic Equations, SIAM, Regional Conference Series, Vol. 1 (1971).

[B5] Bodewig, E., Matrix Calculus, Interscience Publishers, Inc., New York (1959).

[B6] de Boor, C., Mathematical Aspects of Finite Elements in Partial Differential Equations, Academic Press, New York (1974).

[B7] Bramble, J.H. and M.Zlamal, Triangular Elements in the Finite Element Method, Math. of Comp. 24, P. 809-821 (1970).

[B8] Bramble, J.H., J.A.Nitsche and A.H.Schatz, Maximum-Norm Interior Estimates for Ritz-Galerkin Methods, Math. of Comp. 29, p. 677-688 (1975).

[B9] Brandt, A., Multi-level Adaptive Solutions to Boundary Value Problems, Research Rept., Weizmann Institute of Science, Rehovot, Israel IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598

[B10] Broy, M., Numerische Lösung höchdimensionaler Linearer Gleichungssysteme endliecher Netze, Thesis, T.U. München, (1975).

[C1] Collatz, Functional Analysis and Numerical Mathematics, Academic Press, New York (1966).

[C2] Collatz, L., The Numerical Treatment of Differential Equations, Springer-Verlag, New York (1966).

[C3] Courant, R., Variational Method for the Solution of Problems of Equilibrium and Vibrations, Bull. Amer. Math. Soc. 49, P. 1-23 (1943).

[F1] Falk, R.S., Error Estimates for the Approximation of a Class Variational Inequalities, Math. of Comp. 28, P. 963-971 (1974).

[F2] Fedorenko, P.P., The speed of Convergence of One Iterative Process, USSR Computation Math. and Math. Physis, 19, P. 227-235 (1964).

[F3] Frederickson, P.O., Generalized Triangular Splines, Math. Rept. 7, Lakehead University, Canada (1971).

[F4] Frederickson, P.O., Fast Approximate Inversion of Large Sparse Linear System, Math. Rept. 7, Lakehead University, Canada (1975).

[F5] Frederickson, P.O. and W. Eames, Generalized Hardy Inequality, to appear.

[F6] Frederickson, P.O., Affine Analysis, to appear.

[F7] Frederickson, P.O., Fast Solution of Generalized Eigenvalue Problems, to appear.

[H1] Hewitt E. and K. Stromberg, Real and Abstract Analysis, Springer-Verlag, New York (1969).

[H2] Hildebrand, F.B., Advanced Calculus for Applications, Prentice-Hall, Inc., Englewood Cliffs, New Jersey (1976).

[H3] Hoffman K., Analysis In Enclidean Space, Prentice-Hall, Inc., Englewood Cliffs, New Jersey (1975).

[H4] Holand, I., and K. Bell, eds., Finite Element Methods In Stress Analysis, Tapir, Trondheim, Norway (1969).

[K1] Kantorovich and Krylov, Approximate Method of Higher Analysis, Interscience Publishers, Inc., New York (1964).

[L1] Lim, T.H., Higher Dimensional Numerical Quadrature, Thesis, Lakehead University (1974).

[M1] Marchuk, G.I., Methods of Numerical Mathematics, Springer-Verlag, New York, Heidelberg Berlin (1975).

[M2] Martin Schechter, Principles of Functional Analysis, Academic Press, New York (1971).

[M3] Martin, H.S., L^2 Error Bounds for the Rayleigh-Ritz-Galerkin Methods, SIAMJ. Numer. Anal. 8, P. 737-748 (1971).

[N1] Nitsche J.A. and A.H. Schatz, Interior Estimates for Ritz-Galerkin Methods, Math. of Comp. 28, P. 937-958 (1974).

[P1] Prenter, P.M., Splines and Variational Method, John Wiley & Sons, New York (1975).

[S1] Sard, Linear Approximation, Mathematical Survey 9, American Mathematical Society, Providence, Rhode Island (1963).

[S2] Schwartz, D., Linear Operators, Interscience Publishers, Inc., New York (1957).

[S3] Shanks D., Nonlinear Transforms of Divergent and Slowly Convergent Sequences, J. Math. Phys. 34, P. 1-42 (1958).

[S4] Strang, G. and G. Fix, An Analysis of the Finite Element Method, Prentice-Hall, Englewood Cliffs, N.J. (1973).

[T1] Teman, R., Numerical Analysis, D. Reidel Publishing Company, Dordrecht, Holland (1973).

[V1] Varga, R.S., Matrix Iterative Analysis, Prentice-Hall, Inc., Englewood Cliffs, New Jersey (1962).

[V2] Varga, R.S., The Effect of Quadrature Errors in the Numerical Solution of Two-Dimensional Boundary Value problems by Viriational Techniques, Aequationes, Math. 7, P. 36-58 (1972).

[W1] Williamson R.E., R.H. Crowell and H.F. Trotter, Calculus of Vector Functions, Prentice-Hall, Inc., Englewood Cliffs, New Jersey (1968).

[W2] Whiteman, The Mathematics of Finite Elements and Application, Academic Press, New York (1965).

[Y1] Yosida. K, Functional Analysis, Academic Press, New York (1965).

[Y2] Young, D.M., Iterative Solution of Large Linear Systems, Academic Press, New York (1971).

[Z1] Zlamal, M., On the Finite Element Method, Numer. Math., 12 P. 394-409 (1968).